

University of Montana

ScholarWorks at University of Montana

Graduate Student Theses, Dissertations, &
Professional Papers

Graduate School

2003

Reliability generalizations of three alcohol screening measures: The Alcohol Use Disorders Identification Test the CAGE Questionnaire and the Michigan Alcoholism Screening Test

Alan Laramie Shields
The University of Montana

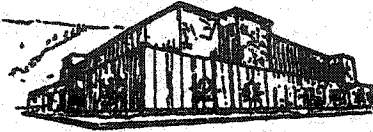
Follow this and additional works at: <https://scholarworks.umt.edu/etd>

Let us know how access to this document benefits you.

Recommended Citation

Shields, Alan Laramie, "Reliability generalizations of three alcohol screening measures: The Alcohol Use Disorders Identification Test the CAGE Questionnaire and the Michigan Alcoholism Screening Test" (2003). *Graduate Student Theses, Dissertations, & Professional Papers*. 9457.
<https://scholarworks.umt.edu/etd/9457>

This Dissertation is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.



**Maureen and Mike
MANSFIELD LIBRARY**

The University of
Montana

The author to reproduce this material in its entirety, provided that this material is used for scholarly purposes and is properly cited in published works and reports, grants permission.

****Please check "Yes" or "No" and provide signature****

Yes, I grant permission

X

No, I do not grant permission

Author's Signature: _____

At-2. S. L. L.

Date: _____

5/13/02

Any copying for commercial purposes or financial gain may be undertaken only with the author's explicit consent.

RELIABILITY GENERALIZATIONS OF THREE ALCOHOL SCREENING

MEASURES:

THE ALCOHOL USE DISORDERS IDENTIFICATION TEST,

THE CAGE QUESTIONNAIRE, AND

THE MICHIGAN ALCOHOLISM SCREENING TEST

By

Alan Laramie Shields, M. A.

The University of Montana, 2000

Presented in partial fulfillment of the requirements

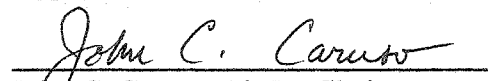
for the degree of

Doctor of Philosophy

The University of Montana

2002

Approved by:


John C. Caruso, Ph.D., Chairperson


Dean, Graduate School

5-15-02

Date

UMI Number: 3089069

UMI[®]

UMI Microform 3089069

Copyright 2003 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Shields, Alan L., M.A., May 2002
Psychology

Reliability Generalizations of Three Alcohol Screening Measures: The Alcohol Use Disorders Identification Test, the Michigan Alcoholism Screening Test, and the CAGE Questionnaire (88 pp.)

Chairperson, John C. Caruso, Ph.D. *JCC*

Objective: The Alcohol Use Disorders Identification Test (AUDIT), the CAGE Questionnaire, and the Michigan Alcoholism Screening Test (MAST) and its two most frequently used derivatives, the Brief MAST (BMAST) and the Short MAST (SMAST) are among the most commonly employed and scientifically investigated alcohol screening measures. While considerable work has been done on the validity of scores generated by these measures, relatively little information is available on their reliability. In order to improve understanding of how scores from these measures perform and give healthcare providers an enhanced empirical foundation to refer to when using them, two goals are outlined: (a) to provide a characterization of the reliability of AUDIT, CAGE, MAST, BMAST, and SMAST scores and (b) to discern methodological or sample characteristics associated with the reliability of those scores.

Method: The meta-analytic technique of reliability generalization was employed. This method integrates all published research using the identified measures and permits far broader analyses of scores generated by the instruments than would be possible in any one study. The present reliability generalizations had two primary goals: (a) to characterize the typical reliability of scores for the measures and (b) to examine factors that may be related to the reliability of those scores.

Results: Only 11% of reviewed studies reported reliability information for the data at hand. The median internal consistency reliabilities for the AUDIT, CAGE, MAST, and SMAST were .81, .74, .85, and .84, respectively. Reliability estimates ranged from .47 to .94 indicating a fair amount of variability across samples and study characteristics. Score reliability and sample age were the two most consistent predictors of score reliability across measures.

Conclusions: Reliability and other psychometric reporting practices within the alcohol screening literature were poor. In fact, the BMAST literature provided too few reliability estimates for the measure to be considered for analysis. Relative to current standards, results suggest that the AUDIT, MAST, and SMAST are generally capable of producing scores with sufficient reliability for most basic research purposes and individuals administering these measures can do so with confidence in these situations. The CAGE demonstrates poor to marginal ability to generate reliable scores. Scores from each measure, however, should not be interpreted in isolation especially when clinically important decisions are based on them.

Table of Contents

Abstract.....	ii
List of Tables.....	v
Introduction.....	1
Alcohol Screening Measures.....	1
Evaluating Measures.....	3
Reliability.....	3
Reliability Defined and Interpreted.....	4
Reliability Indices.....	6
Importance of Reliability.....	8
Reliability as a Property of Scores.....	8
Reporting Reliability.....	9
Reporting Reliability: The State of the Art.....	11
Reliability Generalization.....	14
Reliability Generalization: A Review of Applications.....	14
The Present Study.....	19
General Method.....	19
Selection of Alcohol Screening Measures.....	19
Data Collection.....	20
Criterion Variable.....	21
Predictor Variables.....	21
Data Analysis.....	24
Application 1: The Alcohol Use Disorders Identification Test.....	25
Method: Application 1.....	25
Results: Application 1.....	27
Discussion: Application 1.....	29
Application 2: The CAGE Questionnaire.....	31
Method: Application 2.....	31
Results: Application 2.....	32
Discussion: Application 2.....	34
Application 3: The Michigan Alcoholism Screening Test.....	35
Method: Application 3.....	35
Results: Application 3.....	38
Discussion: Application 3.....	42

General Discussion.....	44
Reliability Reporting.....	44
Reliability Generalization.....	46
Limitations.....	49
Summary.....	51
References.....	53
Appendix A.....	76
Appendix B.....	79
Appendix C.....	80
Appendix D.....	82
Appendix E.....	83

List of Tables

Table 1: *The AUDIT: Descriptive Statistics for Score Reliability and Predictor Variables.*

Table 2: *The AUDIT: Summary Results from Hierarchical Regression Analyses.*

Table 3: *The CAGE Questionnaire: Descriptive Statistics for Score Reliability and Predictor Variables.*

Table 4: *The MAST: Descriptive Statistics for Score Reliability and Predictor Variables.*

Table 5: *The SMAST: Descriptive Statistics for Score Reliability and Predictor Variables.*

Table 6: *Alcohol Screening Measures: Reliability Reporting Practices.*

Table 7: *Alcohol Screening Measures: Descriptive Statistics for Score Reliability.*

Reliability Generalizations of Three Alcohol Screening Measures: The Alcohol Use Disorders Identification Test, The CAGE Questionnaire, and The Michigan Alcohol Screening Test

Results from the National Comorbidity Survey suggest that only 14.5% of individuals with pure alcohol use disorders will utilize treatment services in a given year (Wu, Kouzis, & Leaf, 1999). Furthermore, only 2-5% of lifetime drinkers will ever receive help to deal with problems in their drinking (Ogborne & DeWit, 1999). These findings are disheartening given that over 20% of the general population will suffer from an alcohol use disorder at some point during their life (Kessler et al., 1994). Although relatively cost-effective treatments for alcohol use disorders are available (Holder, Longbaugh, Miller, & Rubonis, 1991) the costs of alcohol abuse remain staggering. In a joint report, the National Institute on Drug Abuse (NIDA) and the National Institute on Alcohol Abuse and Alcoholism (NIAAA) estimated alcohol abuse costs at \$166.5 billion in 1995 alone (NIDA/NIAAA, 1998). Human service groups, researchers, and clinicians are increasingly interested in intensifying treatment opportunities for individuals suffering from alcohol use disorders (Department of Health and Human Services, 1990).

Alcohol Screening Measures

One possible reason why available treatment opportunities are not utilized by those most in need is lack of effective and psychometrically sound alcohol screening measures. In order to more efficiently identify problem drinkers and minimize future personal harm and social cost, strategies are being developed to improve brief alcohol screening measures in a variety of clinical (e.g., Maisto, Carey, Carey, Gordon, & Gleason, 2000) and primary health care settings (e.g., Fleming, 1997; Soderstrom et al.,

1998). Defining alcohol screening measures has proven to be a difficult task and multiple interpretations have been generated. However, NIAAA (1995a) defines and explains screening as:

The skillful use of empirically based procedures for identifying individuals who have alcohol related problems/consequences or who are at high risk for such difficulties. Screening measures are not designed (if for no other reason than their brevity) to explicate the nature and extent of such problems. (p. 17)

These measures are now being validated by gender (e.g., Bradley, Boyd-Wickizer, Powell, & Burman, 1998) and among minority populations (e.g., Cherpitel, 1998). In addition, detailed reviews regarding the goals, methods, use, and performance of brief alcohol screening methods among a variety of populations can be found in the psychological as well as the medical literature (Allen, Cross, Fertig, & Litten, 1998; Bradley, Kivlahan, Bush, McDonell, & Fihn, 2001; Carey & Teitelbaum, 1996; Fiellin, Reid, & O'Connor, 2000; Fleming, 1997; Maisto, Connors, & Allen, 1995; McPherson & Hersch, 2000; Piccinelli et al., 1997; Nilssen & Cone, 1994).

Despite efforts to improve alcohol screening, researchers continue to cite inadequate knowledge about the psychometric performance of alcohol screening measures in various health care settings and among various populations (e.g., Carey & Teitelbaum, 1996; McPherson & Hersch, 2000). The improved understanding of these measures would likely have many benefits. Each benefit, however, should be considered an instrumental step toward the ultimate goal of providing clinicians and test administrators with rules and guidelines by which they can base their decision to use a particular screening instrument. As Allen & Fertig (1995) stated, the "Evaluation of how

alternative [alcohol use] measures fare on validity and reliability, the two primary psychometric characteristics of [the scores produced by] an assessment instrument, can assist in choosing one scale over another” (p. 4).

Evaluating Measures

The psychometric evaluation of scientific observations, measurements, and tests is crucial to scientific progress. Within the behavioral sciences alone, a multitude of methods, procedures, and analytic approaches are available to assist the researcher in making observations (e.g., Cook & Campbell, 1979; Grimm & Yarnold, 1995; Grimm & Yarnold, 2000; Kazdin, 1992; Keppel, 1991). However, as noted above, the two general observation properties of reliability and validity must first be established before a particular assessment can be useful in perpetuating a scientific discipline. Broadly defined, reliability is the extent to which measurements are stable or repeatable and validity is the extent to which a test is measuring what it is purported to measure. There are several types of validity estimates just as there are several types of reliability. The focus of the present inquiry is on the property of reliability, however, these two properties are inextricably linked.

Reliability

Researchers are increasingly documenting insufficient and inaccurate reliability reporting trends across varied psychological measures (cf., Vacha-Haase, Henson, & Caruso, in press) and within psychometric reporting practices generally (Kieffer, Reese, & Thompson, 2001; Meier & Davis, 1990; Whittington, 1998). This implies that the property of reliability, or at least certain aspects of it, may not be well understood. This is unfortunate given the major role it plays in the lives of professional psychologists. In

fact, the role reliability plays is so central to psychology that Li, Rosenthal, & Rubin (1996) stated:

Although psychologists are becoming so specialized and so differentiated that at times we find it difficult to understand one another's literature, there remain a few ideas so fundamental to all of psychology that an understanding of them helps define us as psychologists. One of these fundamental ideas is that of the reliability of measurement. (p. 98)

In order to clarify the often-misunderstood concept of reliability, the following three sections will define and interpret reliability, describe its indices, and, lastly, explain its importance.

Reliability Defined and Interpreted

Reliability is broadly operationalized as the consistent measurement of a particular construct under varying circumstances. In other words, measurements are reliable if they are relatively free from the unsystematic measurement errors caused by the background circumstances in which they were gathered. More technically, classical measurement theory (Lord & Novick, 1968) states that a single observation (observed score; X) is the result of a true score (t) and an error score (e) or,

$$X = t + e. \quad (1)$$

In this equation, (a) t cannot be observed directly but instead is the average score a single respondent would get if he or she was observed an infinite number of times, (b) e represents the segment of observed score that unsystematically fluctuates because of

random error sources (i.e., the background circumstances in which observations were made), and (c) these random error sources can include but are not limited to conditions of measurement administration and respondent characteristics.

When groups of individuals are considered for analysis they are generally analyzed in terms of the variability or dispersion in their scores. The most common measure of dispersion is variance, which represents the degree to which scores tend to differ or vary from the average of the observed scores. The statistic of variance is symbolized as σ^2 with higher numbers indicating greater dispersion. Its most basic formula,

$$\sigma_X^2 = \frac{\sum (X - \mu)^2}{N}, \quad (2)$$

represents the average squared deviation between the observed score (X) and the population mean (μ). Therefore, if the observed score consists of a true score (t) and an error score (e) then observed variance consists of true score variance and error score variance:

$$\sigma_X^2 = \sigma_t^2 + \sigma_e^2 \quad (3)$$

Equation 3 provides enough information to specify classical measurement theory's most basic interpretation of reliability. Specifically,

$$\rho_{XX'} = \frac{\sigma_t^2}{\sigma_X^2} \quad (4)$$

where $\rho_{XX'}$ is the reliability coefficient represented as a ratio of the true score variance to observed score total variance.

The reliability coefficient can range from 0 to 1. The value of 0 is interpreted as the absence of reliability and informs all variability is error. The value of 1 is interpreted as perfect reliability and informs that the observed score variance is fully accounted for by only true score variance; there is no error variance. Measures tend not to produce completely unreliable or completely reliable scores but, instead, generate scores with varying degrees of reliability. Although no rigid rule regarding the acceptability or unacceptability of a particular reliability coefficient exists, current standards for commonly used indices of reliability suggest a minimum score reliability cut-off value of .70 for the early stage of measure development, .80 for basic research purposes, and .90 when important clinical decisions are being made (Nunnally & Bernstein, 1994).

Reliability Indices

Most generally, a reliability index attempts to quantify the degree of random or unsystematic error in scientific observation. Each method of calculating reliability attempts to account for different conditions leading to these unsystematic changes in individual observations that can then affect error variance. While there are likely infinite ways in which random error can be introduced into any single observation and numerous methods for demonstrating consistency across observations, the following four are general classes of reliability estimates found in the empirical literature: (a) inter-rater, (b) test-retest, (c) parallel-forms or alternative forms, and (d) internal consistency.

Unsystematic error can be introduced into observations when different test administrators use the same measure. Inter-rater reliability estimates, therefore, are used

to assess the degree to which different observers give consistent ratings of some characteristic to the same object. The kappa coefficient (κ) is a commonly reported estimate of inter-rater reliability because of its ability to account for the chance agreement possibility between different observers.

The test-retest reliability coefficient is defined as the correlation between the scores generated by the same test administered on two separate occasions. Constructs that are thought to hold constant over time (e.g., personality traits) should produce very similar scores over repeated measurements. Unsystematic differences generated by contrasts in the testing conditions, therefore, represent the error of interest in test-retest reliability methods.

Parallel-forms methods of calculating reliability attempt to understand the error in measurement due to characteristics of the test itself. Specifically, a parallel-forms coefficient is the correlation of scores generated from two separate but equivalent measurements of the same construct.

Lastly, internal consistency methods attempt to delineate those test items that tend not to converge on the conclusions produced by the other test items. In other words, consistency across items is often desirable because it means they are all measuring the same construct. Error in item agreement can be evaluated using internal consistency methods. Internal consistency coefficients are generated from single test administrations of the same measure and they cannot be considered equivalent to test-retest or parallel-forms coefficients. Common methods of internal consistency are the split-half method, Kuder-Richardson method, and coefficient alpha (α).

Importance of Reliability

Attenuation theory (Lord & Novick, 1968) states that the observed correlation between two scores on two different measures is lower than their true score or actual correlation. In other words, measurement error contributes to an attenuated observed relationship between the two variables. For example, if a particular alcohol screen measure generated score X and a measure of psychiatric distress generated score Y , their observed correlation is attenuated or lower than their true score correlation. The difference between an observed correlation and a true score correlation is due to less than perfect reliability of scores produced by the two measures. This unreliability of measures of alcohol use, X , and psychiatric distress, Y , may lead researchers to erroneously conclude that the two variables are unrelated. Inferentially, unreliable scores can contribute to an overall loss of power against a Type II error and lead to errors of inference or a study's inability to detect a significant relationship even when the relationship actually exists. Further, the unreliability of scores generated by X and Y may reduce the observed correlation such that even clinically significant relationships go undetected. Thus, undetected and unsystematic errors in measurement lead to unreliable test scores, and unreliable test scores can significantly interfere with a researcher's power to draw accurate conclusions about the relationships among events, variables, and concepts. This fact has often led to the conclusion that "reliability precedes validity."

Reliability as a Property of Scores

Strube (2000) stated that, "Reliability depends on the conditions of measurement, and there are as many reliabilities for a measure as there are unique uses or conditions of measurement" (p. 32). Thompson (1994) further noted that, "Too few researchers act on

a conscious recognition that *reliability is a characteristic of scores on the data in hand*" (p.839) and that "The subjects themselves impact the reliability of scores, and thus it becomes an oxymoron to speak of 'the reliability of the test' without considering to whom the test was administered or other facets of the measurement protocol" (p. 839). In other words, reliability is not a property of a test but rather a property of scores generated by a test, which, as noted, can be influenced by multiple and various methodological and subject characteristics. While not a novel concept, its importance is severely understated. It is important to the extent that scores from samples with different characteristics (such as differing diagnostic groups, ages, or gender representation) can and will differ with respect to reliability. In this instance, score is operationalized in a broad sense meaning the quantitative summary of observations generated from an assessment device (Messick and the Educational Testing Service, 1995).

Despite this knowledge, it is not unusual to hear both researchers and practitioners refer to a particular test as being more reliable than another. Further, it is not unusual to read in the empirical literature that a particular test has been found to be a reliable measure but offer no reference to the score reliability for the sample under examination (cf., Vacha-Haase et al., in press). While this appears to be only a technicality in semantics, Thompson (1994) succinctly noted that, "Such language is both incorrect and deleterious in its effects on scholarly inquiry, particularly given the pernicious consequences that unconscious paradigmatic beliefs can exact" (p. 839).

Reporting Reliability

With growing awareness that reliability is a property of scores and not tests and that these scores are necessary for interpreting effects sizes, careful guidelines by which

to report these statistics are emerging in the literature. Thompson (1994), as editor of *Educational and Psychological Measurement*, offered specifications for authors interested in submitting articles to the journal. He stated that:

Based on these considerations [reliability as a property of scores], use of wording such as “the reliability of the test” or “the validity of the test” will not be considered acceptable in the journal. Instead, authors should use language such as “the scores in our study had a classical theory test-retest reliability coefficient of X.” (p. 841)

More recently, the Board of Scientific Affairs of the American Psychological Association (APA) assembled the Task Force on Statistical Inference (TFSI). The TFSI was asked to address contentious points surrounding statistical methods in psychology journals. In addition, they were asked to make recommendations regarding the appropriate ways in which statistical results should be reported in the literature. Led by Leland Wilkinson, the TSFI (Wilkinson & the APA/TFSI, 1999) stated:

It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. Interpreting the size of observed effects requires an assessment of the reliability of scores. (p. 596)

The task force is interested in reminding researchers that: (a) reliability is a characteristic of a set of scores and not of tests, (b) the accurate assessment of the reliability coefficient produced by a set of scores will impact the way in which effect sizes are interpreted, and

(c) reliability coefficients from the data being analyzed should be reported each and every time scores of a test are reported (Wilkinson & The APA/TFSL, 1999).

Reliability Reporting: The State of the Art

Despite the importance of reporting reliability estimates in research, even a cursory review of the literature to date suggest that scientists are falling woefully short in terms of actually reporting such data. Vacha-Haase et al. (in press) recently documented this notion in their review of over 20 studies that independently evaluated score reliability reporting for a particular measure or set of measures. While there was great variability in the status of reliability reporting between studies (range = 0% to 71% of published studies in a particular area reported reliability), the majority reported less than 25% of studies in a given literature provided reliability information based on the data in hand (Vacha-Haase et al., in press).

In addition, reviews have been performed on the psychometric reporting practices, including reliability reporting, across entire journal volumes. Willson (1980) reviewed all published quantitative articles (i.e., original research articles involving some data collection and statistical analyses) from *American Educational Research Journal* (*AERJ*) from 1969 to 1978 and stated that 37% of studies reported reliability estimates for the data at hand while 18% reported it from previous research and 55% failed to mention reliability altogether. Further, Meier & Davis (1990) found that sample reliability estimate reporting in *Journal of Counseling Psychology* (*JCP*) quantitative studies decreased then increased from 7% to 3% to 23% from 1967 to 1977 to 1987, respectively.

In a follow-up study, Kieffer et al. (2001) evaluated quantitative studies in both *AERJ* and *JCP* from 1988 to 1997. For *AERJ* articles over this 10-year period, 22.8% reported score reliability while 18.8% and 58.4% cited previously reported reliability estimates or did not mention reliability at all, respectively. For *JCP* articles over this 10-year period, 43.9% reported sample reliability while 15.9% and 40.2% cited previously reported reliability estimates or did not mention reliability at all, respectively. Their general conclusion was that the majority of reviewed articles contained no reports of reliability. Similar results have been reported in similar reviews of other educational and psychological journals (Thompson & Snyder, 1998; Whittington, 1998).

A significant contribution to the larger literature regarding the importance and implications of reliability reporting is the notion of *reliability induction*. Vacha-Haase, Kogan, & Thompson, (2000) coined the term reliability induction in order to draw attention to the fact that researchers commonly reference reliability coefficients from past studies or normative samples and induce that the data in hand, therefore, are equally reliable. By comparing the sample composition and variability of the samples inducing reliability to the cited research samples from which the reliability information was taken, Vacha-Haase, Kogan, & Thompson (2000) examined whether researchers using the Bem Sex Role Inventory and the Rosenberg Self-Esteem Instrument adequately understood the factors that allow this practice to be empirically defensible. They discovered that only 20% of the 270 articles inducing reliability provided adequate information necessary to determine if induction was reasonable. Second, in no case which adequate information was available to compare samples was induction seen as reasonable and often the induction was deemed “starkly implausible” (Vacha-Haase, Kogan, & Thompson, 2000).

Expanding on this notion, Deditius-Island & Caruso (in press) suggest that the omission of reliability coefficients altogether should be considered a special type of reliability induction. They argue that while it is less obvious, failure to provide any reliability information means that the researcher “either (a) does not care about reliability or (b) assumes that the scores will be reliable because they have been so in previous samples, with the former indicating poor research and the latter indicating that a reliability induction has taken place” (Deditius-Island & Caruso, in press). In their reliability generalization of Zuckerman’s Sensation Seeking Scale, Form V, Deditius-Island & Caruso (in press) documented that 214 of 244 (87.7%) articles failed to mention the property of reliability at all. In order to differentiate between the two reliability induction scenarios described by Vacha-Haase, Kogan, & Thompson (2000) and Deditius-Island & Caruso (in press), Shields & Caruso (under review) designated them as Type I and Type II reliability induction, respectively.

Psychometric reporting reviews and reliability induction studies provide convergent and convincing evidence that: (a) researchers typically do not report reliability information in their reports; (b) when researchers report reliability information they frequently extract coefficients from past research of the measure rather than generating coefficients of their own (a practice referred to as reliability induction); and (c) when engaging in reliability induction, researchers tend to do so without an adequate understanding of the assumptions or factors that might make this practice reasonable in a given situation.

Reliability Generalization

Reliability coefficients should be reported on a study-by-study basis because they frequently vary across, among other things, test samples and mode of test administration. This notion anticipates a meta-analytic technique called reliability generalization (Vacha-Haase, 1998). A reliability generalization has two main objectives: (a) To provide a characterization and an index of variability of the reliability of scores for a particular measure and (b) to discern methodological or sample characteristics that are associated with score reliability.

The method of reliability generalization is rooted in another meta-analytic technique explored by Schmidt & Hunter (1977) and Hunter & Schmidt (1990), the validity generalization. The two techniques are similar in that both use published studies as the unit of analysis. What differentiates the techniques is the dependent variable of interest. In validity generalization studies, the criterion is usually a measure of effect size (e.g., correlations) while in reliability generalization, the criterion is an estimate of reliability (e.g., internal consistency). In both, the criterion is to be predicted by select study and sample characteristics.

Reliability Generalization: A Review of Applications

Vacha-Haase (1998) introduced the term “reliability generalization,” described its aims, and offered an account as to why invoking this conceptual meta-analytic model could benefit the broad field of measurement. In addition, she provided a reliability generalization example by investigating the dependent variable of score reliability of the Bem Sex Role Inventory (Vach-Haase, 1998). Her results demonstrated the importance of recognizing reliability as a property of scores rather than measures, as reliability

coefficients tended to widely vary across 87 samples. Further, Vacha-Haase (1998) showed that a variety of sample and method characteristics were important predictors of the reliability of scores. Lastly, she reported that 65.8% of the 628 studies made no mention of reliability at all while only 13.1% of the studies reported adequate reliability information. This was a warning sign for the research community at large to attend to and report important psychometric information, particularly reliability estimates, in their studies.

While Vacha-Haase (1998) is typically the primary reference in reliability generalization studies, it is important to recognize that several other research groups approximated the method prior to her study but not under the title “reliability generalization.” In two similar studies, Schuerger, Tait, & Tavernelli (1982) and Schuerger, Zarrella, & Hotz (1989) used an archival data collection approach and meta-analytic methods to provide estimates of typical test-retest score reliability for eight commercially used self-report personality inventories. They also demonstrated that, in addition to the test-retest interval, other sample characteristics had statistically significant relationships with indices of temporal stability. Several other reliability generalization approximations were performed prior to Vacha-Haase (1998); however, one must look to the marketing and consumer research field to find them. Again, in similar studies Churchill & Peter (1984) and Peter & Churchill (1986) investigated typical score reliability as well as factors influencing those scores in measures used in marketing research. While Peter & Churchill (1986) evaluated score reliability within a broader meta-analytic framework designed to assess a variety of psychometric criteria, Churchill & Peter (1984) specifically addressed score reliability. Both studies provided typical

score reliabilities and empirically demonstrated that those scores are influenced by sample and methodological factors.

Motivated by Churchill & Peter (1984) and Peter & Churchill (1986), Peterson (1994) initiated a far more broad exploration of score reliability within marketing and consumer research and beyond. Whereas the former studies investigated score reliability within marketing journals, the latter evaluated both marketing and psychology journals. By canvassing both of these broad fields, Peterson (1994) accomplished the daunting task of reviewing more than 33,000 articles in order to extract internal consistency estimates (α) and a variety of sample and methodological characteristics thought to influence those estimates. Such a large-scale effort was undertaken in order to characterize actual internal consistency reliability coefficients (i.e., those appearing in the empirical literature) and compare them with contemporary, yet only theoretical, standards of what constitutes adequate score reliability (e.g., Nunnally, 1978). Peterson's (1994) general conclusions were that actual score reliability values compare favorably with the theorized standards and that there were no important relationships between score reliability estimates and research design characteristics.

After Vacha-Haase (1998) coined the phrase "reliability generalization," however, and documented its viability and usefulness, *Educational & Psychological Measurement* dedicated a special section to the method and its related topics (Thompson, 2000). Three reliability generalization studies were highlighted in this special section. First, Yin & Fan (2000) assessed the reliability of Beck Depression Inventory (BDI) scores. Next, Viswesvaran & Ones (2000) evaluated "Big Five Factors" in terms of their reliable estimation across various measures. Lastly, Caruso (2000) applied reliability

generalization to the NEO Personality Scales. By documenting significant variability in score reliability estimates, each of these studies further demonstrated the importance of addressing reliability in terms of scores rather than tests. In addition, each of these studies documented dismal score reliability reporting rates (most notably, Yin & Fan, 2000). These findings were congruent with historical and contemporary reviews of psychometric reporting more generally (Kieffer et al., 2001; Willson, 1980) and support a growing campaign aimed at reminding researchers, journal editors, and professional groups of the necessity of score reliability reporting in their manuscripts, journals, and professional presentations (e.g., Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000).

Two of these studies highlight the broad applicability of the method. Specifically, both Caruso (2000) and Viswesvaran & Ones (2000) evaluated the reliable estimation of personality traits, however, Caruso (2000) evaluated a single personality measure as opposed to a set of personality factors more generally as was done in Viswesvaran & Ones (2000). As reliability generalization is not to be seen as a single, static method, these studies demonstrate the sentiment put forth in Henson & Thompson's (in press) reliability generalization tutorial in which they noted that the only limiting factor in reliability generalization studies, aside from the data reported in the literature base under examination, is the "creativity and insightfulness of the researcher."

Another highlight from the special section was Caruso's (2000) mode of reporting reliability generalization results. He stressed that reliability generalization studies can and should supply service providers with an empirically based and logical set of guidelines for selecting the measure that is most suited to their particular situation.

Accordingly, Caruso (2000) presented his results in a way that provided the clinician with a straightforward list of those situations in which score reliability was too low and use of the measure under study may be questionable.

Following the special section in *Educational and Psychological Measurement* (Thompson, 2000), other reliability generalization studies have been conducted, and include investigations of the children's Junior Eysenck Personality Questionnaire (Caruso & Edwards, 2001), the adult Eysenck Personality Questionnaire (Caruso, Witkiewitz, Belcourt-Dittloff, & Gottlieb, 2001), the Mathematics Anxiety Rating Scale (Capraro, Capraro, & Henson, 2001), the Teacher Efficacy Rating Scale and its related instruments (Henson, Kogan, & Vacha-Haase, 2001), and all the validity (Vacha-Haase, Tani, Kogan, Woodall, & Thompson, 2001) and clinical (Vacha-Haase, Kogan, Tani, & Woodall, 2001) scales of the Minnesota Multiphasic Personality Inventory.

As the importance of score reliability in measurement has gained increased attention, *Educational and Psychological Measurement* has elected to follow up their special section on reliability generalization (Thompson, 2000) by dedicating an entire issue to the method (Thompson, in press). This issue will feature score reliability examinations of the Spielberger State-Trait Anxiety Inventory (Barnes, Har, & Jung, in press), Marlow-Crown Social Desirability Scale, (Beretvas, Meyers, & Leite, in press), Myers-Briggs Type Indicator (Capraro & Capraro, in press), Zuckerman's Sensation Seeking Scales (Deditius-Island & Caruso, in press), Working Alliance Inventory (Hanson, Curry, & Bandalos, in press), Learning Style Inventory (Henson & Hwang, in press), Geriatric Depression Scale (Kieffer & Reese, in press), Coopersmith Self-Esteem Inventory (Lane, White, & Henson, in press), Career Decision-Making Self-Efficacy

Scale (Nilsson, Schmidt, & Meek, in press), Adult Attachment Scale and related instruments (Reese, Kieffer, & Briggs, in press), Alcohol Use Disorders Identification Test (Shields & Caruso, in press), and the Differential Emotions Scale (Youngstrom & Green, in press).

The Present Study

The primary aim of this project was to improve the psychometric understanding and performance of three commonly used alcohol screening measures and to give healthcare providers an enhanced empirical foundation to refer to when using these measures. This was accomplished by applying the meta-analytic technique of reliability generalization to the Alcohol Use Disorders Identification Test, the CAGE Questionnaire, and the Michigan Alcoholism Screening Test and its most commonly used derivatives, the Short MAST and the Brief MAST. More specifically, this project characterized the typical reliability of scores produced by each measure. Further, certain methodological conditions and sample characteristics thought to be associated with score reliability of these measures were evaluated. Because reliability generalization is a meta-analytic method requiring archival data collection, reliability reporting practices more generally were also evaluated.

General Method

Selection of Alcohol Screening Measures

Although a wide variety of alcohol screening measures exist, to be considered for examination in the current project the measure had to meet certain selection criteria. In addition to the eight inclusion criteria for screening instruments outlined in *Assessing*

Alcohol Problems: A Guide for Clinicians and Researchers (NIAAA, 1995a; pages vii-viii), the measures met the following criteria to qualify for study:

1. The measure is in self-report format (either filled out directly by the respondent or given via interview).
2. The measure can be administered to groups as well as individuals.
3. The measure is available for use in a variety of clinical and medical health care settings (e.g., primary care, emergency room, psychiatric patients, and DWI offenders).
4. The measure is brief in nature.
5. The measure is adequately integrated into the larger literature on alcohol screening.
6. The measure's score reliability has not previously been examined.

Using the above criteria, three alcohol screening measures were identified: (a) the Alcohol Use Disorders Identification Test (AUDIT; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993), (b) the CAGE Questionnaire (Mayfield, McLeod, & Hall, 1974), and (c) the Michigan Alcoholism Screening Test (MAST; Selzer, 1971).

Data Collection

Aside from measure selection, the first step in a reliability generalization is to develop a complete list of all studies that have employed the measure to be examined. Using an archival approach, this method allows for a far broader analysis of scores generated by an instrument than would be possible in any one study. In January 2001, a literature search using the American Psychological Association's PsychINFO and the National Library of Medicine's PubMed (Medline) databases was performed for each

measure. PsychINFO lists articles from 1,471 journals in psychology and related disciplines. Further, PubMed (Medline) contains references and author abstracts from more than 4,000 biomedical journals published in the United States and 70 other countries. For each measure, the search was limited to articles published in peer-reviewed journals in English. After a list of articles was generated for each measure and false-positive citations were eliminated, the remaining articles were obtained, read, and categorized by the reliability reporting practice employed.

Criterion Variable

The criterion or dependent variable of interest in reliability generalization studies is generally some numeric index of reliability. While some reliability generalization researchers have performed analyses in which the criterion is made up of different types of reliability estimates (test-retest correlation coefficients and coefficients of internal consistency; e.g., Vacha-Haase, 1998; Capraro et al., 2001), this approach may be “mixing apples and oranges” as these estimates evaluate different sources of measurement error (Dimitrov, in press). Because internal consistency reliability estimates, particularly α (Hogan, Benjamin, Brezinski, 2000), are by far the most commonly reported estimates of reliability, they were identified as the criterion variables.

Predictor Variables

Because the number of potential factors that could influence score reliability is likely infinite, the number of possible predictor or independent variables that can be used in a reliability generalization study is likely infinite. Nevertheless, insufficient sample and method characteristics reporting in the literature of the measure of interest will likely put restrictions on both the number and type of predictor variables that can be coded

(Henson & Thompson, in press). With these limitations in mind and based on classical test theory and the larger alcohol screening literature, six predictor variables were selected for investigation in the present applications, namely, score variability, age, gender, ethnicity, test version, and sample type. It should be noted that, due to insufficient reporting in the literature, not all identified predictor variables were available for investigation for each application. Deviations from this list are reported in the method section of the relevant application.

Score Variability. Score variability can demonstrate tremendous predictive power as regards score reliability (e.g., Caruso et al., 2001). This is consistent with classical test theory (e.g., Lord & Novick, 1968) which can show that if error variance remains constant and observed score variance increases, then true score variance must increase to exactly the same extent (cf., Henson, 2001). Therefore, the ratio of true score variance to total variance, the reliability, also must increase. Without its inclusion as a predictor variable, variability in score reliability that is really due to increases in observed score variability may be erroneously attributed to other predictor variables. For this reason, the standard deviation of measure scores will be included as a predictor variable.

Age. Research has demonstrated that alcohol use disorders are prevalent and an important source of social and health care related problems and mortality in both the elderly (Adams, Magruder-Habib, Trued, & Broome, 1992; Adams, Yuan, Barboriak, & Rimm, 1993; Smith-Black, Rabins, & McGuire, 1998) and children (Reinherz, Giaconia, Rose, Lefkowitz, & Pakiz, 1993). Brief alcohol screening measures have been used in studies of older adults (e.g., Joseph, Rasmussen, Ganzini, & Atkinson, 1997) and are becoming more widely utilized in studies of younger populations (e.g., Chung et al.,

2000; Thom, Herring, & Judd, 1999) since most adolescent specific alcohol screening measures are long and relatively cumbersome. Given the wide age-range in populations that have been administered the identified screening measures, it is important to determine which measures have a greater likelihood of generating reliable scores among different age groups. To this end, average sample age will be used as a predictor variable within each of the reliability generalizations.

Gender. A recent review of the literature pointed to the paucity of gender-specific analyses found in alcohol screening validation studies (Bradley et al., 1998). A meta-analysis performed in the same report found that alcohol screening measures can perform differently among women and men (Bradley et al., 1998). Therefore, gender is included as a predictor variable and coded as proportion of the sample that is male.

Ethnicity. A gap exists in our knowledge of the performance of alcohol screening measures among minority populations (Steinbauer, Cantor, Holzer, & Volk, 1998). Researchers have made great strides in addressing this problem (e.g., Cherpitel, 1998; Volk, Steinbauer, Cantor, & Holzer, 1997; Cherpitel & Clark, 1995), but little is known about the overall reliability of scores on alcohol screening measures among these populations. Therefore, ethnicity will be used as predictor variable to begin drawing conclusions about the usefulness of these screening measures across ethnic minority groups. This variable will be coded as the proportion of the sample that is Caucasian.

Test Version. The alcohol screening measures selected for analyses are popular assessment tools and have been in existence for many years. While each measure is most frequently used as reported in the present study, alternate versions have been introduced into the larger alcohol screening literature over time. Because altering test items, test

item order, or test item length can introduce unsystematic error into the ultimate measurement scores (e.g., Caruso, 2000; Vacha-Haase, 1998), the version of the alcohol screening measure used will be considered as a predictor variable.

Sample Type. Alcohol use disorders and alcohol use are common problems in a wide variety of psychiatric and health care environments (Gallanter, Castaneda, & Ferman, 1988; Moore et al., 1989). These settings can provide vital opportunities for health care professionals to intervene with alcohol problems at potentially early stages of development. However, the empirical literature suggests that alcohol screening is not routinely applied and alcohol use disorders go under-identified (Duszyński, Nicto, & Valente, 1995; Wolford et al., 1999). One possible explanation for this lack of utilization is that health care professionals are not aware of which measures to use with certain populations. In order to better understand how each of the identified measures performs differing sample populations, sample type will also be used as a predictor variable. It was dummy-coded as clinical (1) and non-clinical (2) sample type. Clinical sample type represents any sample that was assessed in a mental or medical health treatment location (e.g., primary care, emergency room, psychiatric facility, substance use treatment center).

Data Analysis

Traditional descriptive statistics can accomplish the first goal of reliability generalization, to characterize the typical reliability of scores across samples. Data analytic methods within the general linear model are often employed to accomplish the second goal of reliability generalization, to investigate associations between study and sample characteristics and score reliability. In this way, variation in reliability estimates can be systematically explored for their controlling factors. For the present applications,

multiple regression was employed to determine the relationship between the predictor variables and score reliability for each measure. Because score variability is directly related to score reliability under certain basic assumptions, a hierarchical analysis with score variability entered in Model 1, and the other predictor variables entered as a block in Model 2 was employed.

The possibility of using multiple regression is ruled out if the literature contains too few score reliability estimates or omits other important sample or method information. In this case, bivariate correlational analyses can be employed to explore the relationship between the variables of score reliability and the identified continuous predictor variables (e.g., score variability, age, and gender) and point biserial correlation is used to relate score reliability to the dichotomous variable of sample type. These latter techniques were necessary for both the CAGE and the MAST applications.

Application 1: The Alcohol Use Disorders Identification Test

Method: Application 1

The Alcohol Use Disorders Identification Test. The Alcohol Use Disorders Identification Test (AUDIT; Saunders et al., 1993) is a self-report screening measure designed to assess current drinking patterns. It can be given as a pencil and paper self-report or in an interview format by a health professional or paraprofessional. The measure requires about 5 minutes to administer. Each of the measures 10 items is scored on a 0 to 4 ordinal scale creating a range of 0 to 40 where higher scores represent an increased probability of alcohol problems (see Appendix A). A cut-off score of 8 has shown good sensitivity and adequate specificity of alcohol problems (Conigrave, Hall, &

Saunders, 1995), however, alternate cut-off scores may improve AUDIT performance within certain populations (e.g., females; Cherpitel & Borges, 2000).

Based on responses from nearly 2,000 participants from six countries, AUDIT items were originally selected to reflect three factors thought to be closely associated with harmful and hazardous drinking patterns. These factors include: (a) alcohol consumption, (b) alcohol dependence symptoms, and (c) adverse consequences of alcohol use (Saunders et al., 1993). To date, however, there is mixed evidence regarding the actual factor structure of the AUDIT (e.g., Skipsey, Burleson, & Kranzler, 1997; Maisto, Conigliaro, McNeil, Kraemer, & Kelley, 2000) and it has been shown that the factor structure can differ between samples (Shields, Guttmanova, & Caruso, under review). Thus, while it is possible to score the AUDIT for its intended three-factor model, it is primarily used as a single-factor measure generating one overall test score. In addition, as very few studies presented reliability coefficients for these subscales, the total score will be the focus.

Data Collection. Under the guidelines described above and using the search term, “Alcohol Use Disorders Identification Test” (Note: the search terms “AUDIT” and “Alcohol Use Screening Measures” did not reveal any additional studies), a total of 115 studies employing the AUDIT were found. This indicates thorough penetration of the measure into the alcohol literature since its introduction about a decade ago. Eleven articles (9.6%) were either letter’s to the editor, review papers, or other non-empirical reports. This left a total of 104 original empirical articles that could be considered for analysis.

Results: Application 1

Reliability Reporting: The AUDIT. Nearly 80% of AUDIT studies reported on reliability via induction methods. Six articles (5.8%) mentioned the property of reliability but provided no data in support of their claim. Five (4.8%) reported reliability coefficients from other studies employing the AUDIT, but not their own scores. Therefore, by citing previous research regarding the score reliability of the AUDIT, authors of a total of 11 (10.6%) studies practiced Type I reliability induction. By failing to mention reliability at all, the majority ($n = 70$ or 67.3%) of the AUDIT study authors practiced Type II reliability induction.

Twenty-three studies (22.1%) reported reliability coefficients based on the data generated for the report. Of these, however, six insufficiently reported other sample characteristic data or total score reliability coefficients, thereby precluding their use here. Because some studies reported more than one coefficient, the remaining 17 studies (16.3%) generated a usable set of 24 samples. In every instance reliability was reported as α . The sample characteristics of score variability, age, gender, and sample type were also collected. Insufficient data was available to include ethnicity as a predictor variable. Age was coded as average age of the sample. Gender was coded as proportion male. Sample group was coded dichotomously as either a patient (1) or non-patient (0) sample. Score variability was recorded as the standard deviation of the AUDIT scores.

Table 1 provides descriptive statistics for AUDIT score reliability and each of the predictor variables across studies. Of particular interest are the statistics for score reliability. The unweighted mean (.79) and median (.81) values indicate that the AUDIT generally produces scores with adequate reliability. The median is slightly higher than

the mean due to the modest negative skew of the distribution of reliability coefficients.

The range indicates that there is a substantial degree of variability in the reliability of AUDIT scores.

Table 1

The AUDIT: Descriptive Statistics for Score Reliability and Predictor Variables (n = 24).

Variable	Mean	Median	SD	Range
Score reliability (α)	.79	.81	.10	.59 - .94
SD of AUDIT scores	5.3	5.3	2.1	2.0 - 11.9
Age	36.1	37.1	9.1	16.5 - 53.0
Proportion male	.62	.64	.25	.0 - 1.0
Sample type	.58	--	.50	.0 - 1.0

Note. Sample type was coded as patient (1) and non-patient (0).

Table 2 presents the results of the multiple regression analyses. As expected, score variability predicted a statistically significant and large amount of variance in score reliability: $R^2 = .68$ ($F(1, 22) = 47.01, p < .001$), Adjusted $R^2 = .67$ in Model 1. When the other predictor variables were entered in Model 2, none of them was statistically significant: $R^2 = .73$ [R^2 change = .045 (F change (3, 19) = 1.03, $p = .401$, Adjusted $R^2 = .67$]]. Because the power of the tests in Model 2 were low due to the small sample size, a post-hoc power analysis was conducted and indicated that the power to detect the effect of each predictor, in the order indicated in Table 2, was .999, .347, .081, and .059. In addition, a more liberal approach was taken that individually entered each of the predictor variables originally entered simultaneously in Model 2, and still none was

statistically significant. This provides some evidence that none of the predictor variables examined in the present study is able to account for variance in AUDIT score reliability after score variability is taken into account. Note, however, that the value of Pearson r (not controlling for the relationship between score variability and score reliability) between age and score reliability (-.41, 16.8% shared variance) was statistically significant ($p=.045$). Sample type explained 9% of the variance in reliability coefficients.

Table 2

The AUDIT: Summary Results from Hierarchical Regression Analyses.

	B	β	t	p	Pearson r
Model 1					
Score variability	.040	.825	6.86	<.001	.83
Model 2					
Score variability	.036	.747	5.38	<.001	.83
Age	-.002	-.215	-1.65	.115	-.41
Proportion male	.028	.066	.54	.590	.13
Sample Type	.008	.040	.29	.770	.30

Note. The statistical significance tests presented above apply to the regression coefficients, not the Pearson r values. The Pearson r between age and score reliability was statistically significant, $p=.045$.

Discussion: Application 1

The descriptive statistics indicated that scores on the AUDIT, despite being limited by a relatively small number of items (10), are generally reliable. Nunnally & Bernstein (1994) recommend a minimum score reliability cut-off value of .80 for

measurement in basic research. While not spectacular, the values of .79 and .81 for the mean and median score reliability indicate adequate proportions of true score variance for this type of work, however, the range indicates that at times the AUDIT does not produce reliable scores. Nevertheless, the interquartile range of .72 to .87 indicates that AUDIT scores generally have adequate reliability. Some practitioners demand higher score reliability for individual assessment and higher stakes outcomes and a score reliability cut-off of .90 is recommended when important clinical decisions are being made (Nunnally & Bernstein, 1994). Therefore, when employed in situations where personal and social costs are significant, the AUDIT should be used with caution.

As expected, the results of the regression analyses clearly indicated that score variability was most predictive of score reliability. It is possible that the other predictor variables account for the differing score variabilities and, therefore, indirectly impact score reliability. To test this, a regression analyses using age, gender, and sample type as predictors explained a combined 25% of variance in the criterion of score variability. This was not statistically significant (adjusted $R^2 = .139$; $F(3, 20) = 2.24$, $p = .115$). Sample type was the only predictor that accounted for a statistically significant amount of the variance accounted for in score variability ($B = 1.80$, $\beta = .427$, $t = 2.15(3,23)$, $p = .04$). Because variation in score variability was not appreciably influenced by age, gender, and sample type, their possible indirect effect on variability in score reliability is unlikely.

Although somewhat tempered by the low power of the tests due to small sample size, the fact that no other predictor examined here accounted for variance in score reliability over and above that accounted for by score variability can tentatively be

considered a positive characteristic of the AUDIT; it produced equally reliable scores across ages, genders, and sample types, once score variability is accounted for. Future researchers using the AUDIT, however, may want to consider the possibility that non-patient samples may produce scores with low variability thereby indirectly decreasing score reliability.

Application 2: The CAGE Questionnaire

Method: Application 2

The Cage Questionnaire. The CAGE Questionnaire (Mayfield et al., 1974) is one of the most extensively studied self-report alcohol screening measures (Maisto et al., 1995). This measure has been used for the early identification of past and current alcohol use disorders. It is a brief and easily administered, scored, and interpreted alcohol screening measure consisting of a four item mnemonic: 1) Have you ever felt you should *Cut* down on your drinking? 2) Have people *Annoyed* you by criticizing your drinking? 3) Have you ever felt bad or *Guilty* about your drinking? And 4) Have you ever had a drink first thing in the morning to steady your nerves or to get rid of a hangover (*Eye-opener*)? Items are scored either 0 for negative response and 1 for positive response, with a higher score indicative of alcohol problems (see Appendix B). Typically, a total score of 2 or greater is deemed clinically significant (i.e., an increased likelihood the respondent will meet DSM-IV criteria for an alcohol use disorder).

The CAGE targets adult and adolescent (over the age of 16 years old) populations and can be administered as a pencil and paper self-report or in an interview format by a health professional or technician. Scoring is simple and instantaneous. There is no fee for use. A literature search revealed 379 published articles on the CAGE questionnaire.

Data Collection. Under the guidelines described above and using the search terms, “CAGE AND alcohol,” “CAGE alcohol questionnaire,” “CAGE AND alcohol use,” and “CAGE AND alcohol screening,” a total of 316 studies employing the CAGE was found. Of these 316 articles, 11 could not be located. Of the remaining 305, 46 (15.1%) were either letter’s to the editor, review papers, or other non-empirical reports. This left a total of 259 original empirical articles that employed the CAGE and could be considered for analysis.

Results: Application 2

Reliability Reporting: The CAGE Questionnaire. Over 90% of the CAGE studies reported on reliability via induction methods. Thirteen articles referenced the property of reliability but failed to provide supporting data and 227 made no mention of reliability at all. Therefore, by citing previous research regarding the score reliability of the CAGE, authors of 5.0% of studies practiced Type I reliability induction. By failing to mention reliability at all, the majority (87.6%) of authors of CAGE studies practiced Type II reliability induction. Authors of a total of 19 studies employing the CAGE (7.3%) reported reliability coefficients quantified from the data at hand.

Of the 19 CAGE articles that included score reliability information, several had to be eliminated from the analysis. Two were eliminated because they evaluated significantly modified versions of the CAGE. In addition, three studies evaluated inter-rater reliability and another evaluated only test-retest reliability. These were eliminated in order to avoid problems associated with generalizing across different estimates of reliability. A total of 13 studies reported internal consistency reliability (all coefficient α). The 13 studies reporting usable reliability information generated a data set of 22

samples. The sample characteristics of age, gender, sample type, and score variability were also collected. Age was coded as average age of the sample. Gender was coded as proportion male. Sample group was coded dichotomously as either a clinical (1) or non-clinical (0) sample. Score variability was recorded as the standard deviation of the CAGE scores.

Table 3 provides descriptive statistics for all variables examined in the CAGE reliability generalization. Reliability coefficients ranged from .52 to .90 (range = .38) indicating substantial variability in the reliability of CAGE scores across samples. The unweighted mean (.74) and median (.74) values indicate that the CAGE demonstrates poor to marginal ability to produces scores with adequate reliability.

Table 3

The CAGE Questionnaire: Descriptive Statistics for Score Reliability and Predictor Variables.

Variable	Mean	Median	SD	Range	<i>n</i>
Score reliability (α)	.74	.74	.09	.52 - .90	22
SD of CAGE scores	1.13	1.21	.21	.80 – 1.30	7
Average age	32.8	34.4	7.68	17.4 – 44.9	20
Proportion male	.51	.58	.26	0 - .80	14
Proportion white	.61	.73	.31	.16 - .98	10
Sample type (% patient)	.73	--	.46	0 - 1.0	22

Note. Sample type was coded as patient (1) and non-patient (0).

Bivariate correlations examining the association between score reliability and sample age, gender, and ethnicity were performed. The Pearson r was not statistically significant between score reliability and gender ($r = -.16, p = .585, n = 14$), approached statistical significance between score reliability and ethnicity ($r = -.61, p = .06, n = 10$), and was statistically significant between score reliability and age ($r = .48, p = .03, n = 20$). The point biserial correlation between score reliability and sample type approached statistical significance ($r_{pb}(20) = .41, p = .055$).

Discussion: Application 2

Results indicate that the CAGE, relative to current standards, demonstrates poor to marginal ability to generate reliable scores. This may, at least in part, be due to its small number of items (4). A score reliability cut-off level of .80 has been recommended for most basic research purposes (Nunnally & Bernstein, 1994). The value of .74 for both the mean and median score reliability indicate inadequate proportions of true score variance and make questionable its use in basic research. Specifically, 72.7% ($n = 16$) of the obtained samples failed to meet this minimal requirement. No CAGE studies reported score reliability estimates above the cut-off value of .90 recommended by Nunnally & Bernstein (1994) for situations where personal and social costs are significant (e.g., diagnosis). Therefore, interpreting isolated CAGE scores in high-stakes situations should be avoided altogether.

Purpose number two of reliability generalization is to explore sample and method characteristics associated with score reliability. Limiting the use of more sophisticated analyses typically used in reliability generalization studies and tempering the results presented here was the insufficient score reliability and sample characteristics reporting

from the larger CAGE literature. However, bivariate correlation analyses illuminated a statistically significant relationship between CAGE score reliability and sample age. Specifically, results suggest that CAGE score reliability is positively related to age. Sample age ranged from 17 to 44 ($M = 32.8$, $SD = 7.68$) years with older CAGE respondents producing more reliable scores than younger ones. This may begin to explain some findings regarding CAGE score validity. Research shows that CAGE scores have relatively low sensitivity and positive predictive value, particularly relevant indices of validity for screening tools, among college students (Heck & Lichtenberg, 1990; Nystrom, Perasalo, Salaspuro, 1993; Smith, Collins, Kreisberg, Volpicelli, & Alterman, 1987). Of the 22 samples used in the present analysis, the samples made up entirely of undergraduates produced the three lowest score reliability estimates (.63, .58, .52).

Application 3: The Michigan Alcoholism Screening Test

Method: Application 3

*The Michigan Alcoholism Screening Test*¹. The Michigan Alcoholism Screening Test (MAST; Selzer, 1971; see Appendix C) was the first published self-report alcohol screening instrument. Because of this, the measure is internationally known and has been validated in dozens of studies across such diverse populations including, alcohol use patients, drug use patients, psychiatric patients, and a wide variety of medical patients

¹Numerous derivations of the original MAST have appeared in the larger alcohol use screening literature (e.g., the Malmo Modification of the MAST; Nystrom et al., 1993). For the present application, however, only the MAST and its two most commonly used derivatives, the Brief MAST and the Short MAST, are evaluated.

(cf., Storgaard, Nielsen, & Gludd, 1994). Like the CAGE, the MAST was developed for the early identification of alcohol use disorders among clinical populations.

The measure's 25 items converge on symptoms of problematic drinking and negative consequences associated with the use of alcohol. Items are answered in a "yes" or "no" fashion but differentially weighted as 0, 1, 2, or 5 yielding a total score from 0 to 53. The first item, which indicates drinking status of the respondent, is not scored. Because of this, the MAST is often referred to as a 24 item measure. It takes about 15 minutes to complete and scores above 5 are usually indicative of the presence of respondent alcohol problems. However, higher cut-off scores have been used to improve the overall specificity of the measure (e.g., Connors, O'Farrell, Cutter, Logan, & Thompson, 1987; Ross, Gavin, & Skinner, 1990). The MAST can be administered as a pencil and paper measure, via computer, or delivered as an interview. Intensive training is not required for administration, scoring, or interpretation.

Because the MAST is a commonly used measure in a variety of clinical settings, professionals became concerned that the measure was too long and developed shorter versions. The most commonly used and researched of these are the Brief MAST (BMAST; Pokorny, Miller, & Kaplan, 1972; see Appendix D) and the Short MAST (SMAST; Selzer, Vinokur, & Van Rooijan, 1975; see Appendix E). Each of these, like the original MAST, is designed to screen for alcohol use disorders and associated problems.

The Brief Michigan Alcoholism Screening Test. The BMAST is composed of 10 of the original 25 MAST items thought to be central to diagnosing alcoholism. Each item was selected based on its performance in discriminating alcoholics from non-alcoholics

in the original MAST study (Selzer, 1971). In the original BMAST validation study, Pokorny et al. (1972) administered the original MAST, extracted the 10 BMAST items and found correlation coefficients of .95 and .96 between the two forms on 60 patients being treated for alcohol problems and 62 patients presenting with a variety of psychiatric disorders but no known alcohol problems, respectively. As in the original MAST, the 10 items on the BMAST are answered “yes” or “no” and retain the original measure’s weighted scoring scale. Total scores can range from 0 to 29. Typically, a score of 0 indicates no alcohol problems, 1 to 5 indicates potential alcohol problems, and a score of 6 or more indicates the presence of alcohol problems.

The Short Michigan Alcoholism Screening Test. The SMAST is made up of 13 items from the original MAST and each item is also answered as “yes” or “no.” SMAST originators (Selzer et al., 1975) were interested in devising a test that was shorter than the original MAST and easier to score. To this end, any item that was endorsed as “alcoholism-indicating” was simply given one point and the original weighted scoring system was abandoned. The SMAST items were obtained by administering the MAST to 501 male drivers (102 randomly renewing licenses, 171 following a traffic violation, 129 undergoing inpatient alcohol treatment, 99 undergoing outpatient alcohol treatment) and extracting items that “significantly improved the prediction of the dependent variable, which was alcoholic or nonalcoholic” (Selzer et al., 1975, p. 123) via stepwise regression (Selzer et al., 1975). The two forms were found to correlate very highly over the entire sample ($r = .97$). For the SMAST, scores from 0 to 1 indicate no alcohol problem, a score of 2 indicates potential alcohol problems, and scores over 3 indicates the presence of alcohol problems. Follow-up studies have been performed evaluating the SMAST

with Selzer's (1971) original weighting system. In this case, scores can range from 0 to 33 and scores above 5 indicate positive screening results (e.g., Barry & Fleming, 1993).

Data Collection. Under the guidelines described above and using the search terms, "Michigan Alcoholism Screening Test," "MAST," "Short Michigan Alcoholism Screening Test," "SMAST," "Brief Michigan Alcoholism Screening Test," and "BMAST," a total of 491 studies employing these measures were found. Of these 491 articles, 14 could not be located. Of the remaining 477, 31 (6.5%) were either letter's to the editor, review papers, or other non-empirical reports. This left a total of 446 original empirical articles that employed the MAST and its most commonly employed derivatives, the SMAST and BMAST.

Results: Application 3

Reliability Reporting: The MAST, SMAST, and BMAST. A total of 395 or 88.8% of the 446 studies employing the MAST, SMAST, and BMAST used reliability induction to report score reliability. Of these, 91 studies referenced the property of reliability but failed to provide the data necessary to support that claim while 306 studies made no mention of reliability at all. By citing previous research regarding the score reliability of the MAST, authors of a total of 20% of studies practiced Type I reliability induction. By failing to mention reliability at all, 68.6% of the MAST, SMAST, and BMAST study authors practiced Type II reliability induction. A total 49 studies employing these measures (11.0%) contained reliability coefficients quantified from the data at hand.

Of the 49 articles that included score reliability information, several had to be eliminated from the analysis. In order to avoid problems associated with generalizing across different estimates of reliability, three were eliminated because they evaluated

inter-rater reliability and test-retest reliability. Another five studies were eliminated because the reliability coefficients they provided were reported in previous research on the same sample. Their inclusion would have violated the assumption of independence of samples which suggests that each experimental participant (in this case, a sample) contribute data only once in each analysis. A total of 41 studies reported internal consistency reliability ($n = 38$ coefficient α , $n = 2$ split-half, $n = 1$ Kuder-Richardson).

It was initially proposed that the MAST and its commonly used derivatives would be jointly examined via multiple regression. The design included test version as a predictor variable in model II (see data analysis, above). However, insufficient reliability and sample and method characteristics precluded this type of analysis. Therefore, separate correlational analyses were run for the MAST and SMAST. Insufficient reliability information provided for the BMAST precluded statistical analysis altogether. For the MAST reliability generalization, 28 studies produced a data set of 33 samples. For the SMAST reliability generalization, 15 studies produced a data set of 21 samples. The sample characteristics of age, gender, sample type, and score variability were also collected. As before, age was coded as average age of the sample. Gender was coded as proportion male. Sample group was coded dichotomously as either a clinical (1) or non-clinical (0) sample. Score variability was recorded as the standard deviation of the MAST or SMAST scores.

Table 4 provides descriptive statistics for all variables examined in the MAST reliability generalization. Reliability coefficients ranged from .47 to .91 (range = .44) indicating substantial variability in the reliability of MAST scores across samples. The unweighted mean (.82) and median (.85) values indicate that the MAST demonstrates

moderate to good ability to produce scores with adequate reliability. The median is slightly higher than the mean due to the negative skew of the distribution of reliability coefficients.

Table 4

The MAST: Descriptive Statistics for Score Reliability and Predictor Variables.

Variable	Mean	Median	SD	Range	<i>n</i>
Score reliability (α)	.82	.85	.10	.47 – .91	33
SD of MAST scores	9.29	10.2	5.08	1.36 – 19.22	7
Average age	32.2	33.0	6.76	19.6 – 44.0	26
Proportion male	.67	.76	.30	.00 – 1.0	33
Proportion white	.64	.79	.37	.00 – .98	18
Sample type (% patient)	.70	--	.47	.00 – 1.0	33

Note. Sample type was coded as patient (1) and non-patient (0).

Table 5 provides descriptive statistics for all variables examined in the SMAST reliability generalization. Results are similar to those of the longer MAST, especially those pertaining to score reliability. The unweighted mean (.79) and median (.84) values indicate that the SMAST is typically capable of generating scores with adequate reliability. There was also substantial variability in the reliability of SMAST scores across samples as noted by the range of .57 to .93 (range = .44). Also like the MAST, the median is slightly higher than the mean due to the negative skew of the distribution of reliability coefficients.

Table 5

The SMAST: Descriptive Statistics for Score Reliability and Predictor Variables.

Variable	Mean	Median	SD	Range	<i>n</i>
Score reliability (α)	.79	.84	.10	.57 – .93	21
SD of SMAST scores	3.89	3.15	2.38	2.20 – 8.69	6
Average age	34.6	33.6	6.06	21.6 – 44.6	19
Proportion male	.70	.69	.22	.35 – 1.0	20
Proportion white	.65	.69	.32	.00 – .98	15
Sample type (% patient)	.81	--	.40	.00 – 1.0	21

Note. Sample type was coded as patient (1) and non-patient (0).

Bivariate correlations examining the association between score reliability and sample age, gender, and ethnicity were performed. The Pearson r was positive and statistically significant between MAST score reliability and MAST score variability ($r = .66, p = .002, n = 20$), sample age ($r = .43, p = .03, n = 26$), and sample gender ($r = .37, p = .04, n = 33$). There was not a statistically significant relationship between MAST score reliability and sample ethnicity ($r = .28, p = .257, n = 18$). There was also a large mean difference in score reliability between the clinical ($M = .86, SD = .06$) and non-clinical ($M = .72, SD = .10$) groups. The point biserial correlation between MAST score reliability and sample type was positive and statistically significant ($r_{pb}(33) = .64, p < .0005$).

Correlational analysis for SMAST score reliability and the predictor variables did not yield any trends. The Pearson r was not statistically significant between SMAST

score reliability and score variability ($r = .54, p = .266, n = 6$), sample age ($r = -.09, p = .709, n = 19$), sample gender ($r = .09, p = .716, n = 20$), and sample ethnicity ($r = -.22, p = .443, n = 15$). The point biserial correlation between SMAST score reliability and sample type was also not statistically significant ($r_{pb}(33) = .004, p = .986$).

Discussion: Application 3

In terms of typical score reliability, it is not surprising that the longer MAST produced more reliable scores than did the SMAST. Relative to current standards (cf., Nunnally & Bernstein, 1994), however, both measures appear to be capable of generating scores with sufficient reliability for most basic research purposes. Typical score reliabilities along with the negatively skewed distributions around the means of .82 and .79 for the MAST and SMAST, respectively, indicate adequate quantities of true score variance for general research. Nevertheless, only 14% ($n = 3$) of SMAST and 21% ($n = 7$) of MAST studies found reliability coefficients above the recommended cut-off of .90 for use in clinically relevant situations (Nunnally & Bernstein, 1994). This suggests that researchers and clinicians should exercise caution when using either of these measures in high-stakes situations or when making clinically important decisions.

Consistent with classic test theory (Lord & Novick, 1968), there was a statistically significant relationship between MAST score reliability and score variability. This was not true for SMAST scores. However, the small sample size ($n = 6$) used in that analysis limited its power to detect a statistically significant correlation. Based on results from the present AUDIT application and previous research (Caruso et al., 2001), there is good reason to speculate that an increased sample size would generate the statistical power to detect that relationship.

In general, the results suggest that MAST score reliability is more influenced by sample and method characteristics than SMAST score reliability. There was a statistically significant relationship between MAST score reliability and sample age but not between SMAST score reliability and age. Sample age ranged from 19 to 44 years ($M = 32.2$, $SD = 6.76$) with older MAST respondents producing more reliable scores than younger ones. In addition, there was a statistically significant relationship between MAST score reliability and gender suggesting that men tend to produce more reliable scores on the measure than women. This relationship was not observed in the SMAST data. Lastly, the statistically significant point-biserial correlation in score reliability suggests the MAST generates more reliable scores in clinical samples versus non-clinical samples. Again, this relationship was not observed in the SMAST data. Despite generating similar age, gender, and sample type statistics to those generated by the MAST studies (see Tables 4 and 5), SMAST score reliability was unaffected by this variable.

It makes sense that MAST scores tend to be more reliable when gathered from clinical samples made up of older males, as this is the type of sample in which the measure was primarily intended for use. While adequate score reliability in this population seems generally attainable with the MAST, the decreased reliability among women is unfortunate. Of the 10 MAST studies with primarily female samples, 8 (80%) reported reliability coefficients below the .80 recommended cut-off for basic research. Alternatively, only 4 of 23 (17.4%) primarily male samples failed to meet this minimum. As gender-specific analyses are appreciably under-represented in the larger alcohol screening literature (Bradley et al., 1998), these findings may contribute toward devising

alcohol screening tools more appropriate for females. Because the reliability of SMAST scores appear to be more resistant to the influence of gender, it may be a more appropriate alcohol screen for women than is the MAST.

General Discussion

Reliability Reporting

Table 6 demonstrates the dramatic lack of reliability reporting in the alcohol screening literature reviewed in the present research. Only 91 of the 809 studies employing the CAGE, AUDIT, MAST, SMAST, and BMAST reported reliability information based on the data in hand while the vast majority of studies used reliability induction. The most glaring demonstration of failure to consider reliability occurred within the CAGE literature. As noted, almost 93% of all CAGE studies reviewed in the present research met criteria for either Type I or Type II reliability induction, however, several other observations are noteworthy. Of the 13 studies utilizing Type I reliability induction, only two cited research that actually contained reliability information while three failed to make any reference to their claim at all (i.e., the authors simply stated the CAGE was a reliable instrument). More importantly, eight made inaccurate references regarding reliability. That is, they cited work that neither provided any numerical index for score reliability nor addressed the concept more broadly. For example, Rydon, Redman, Sanson-Fisher, & Reid (1992) stated that they selected the CAGE questionnaire because it has "been shown to have considerable reliability and validity in detecting clinical alcoholics" (p. 198). However, they reference studies by Ewing (1984) and Mayfield et al. (1974), which do not evaluate CAGE, score reliability nor make any mention of reliability in general. Furthermore, Reynaud, Malet, Facy, & Glanddier

(2000) state, "The reliability of the CAGE is well recognized, particularly when used in hospital settings" (p. 1059). In addition to citing the Ewing (1984) and Mayfield et al. (1974) studies, these researchers reference studies by Steinweg & Worth (1993) and Bush, Shaw, Cleary, DelBanco, & Aronson (1987) yet none addresses the topic of reliability. Therefore, of the 13 CAGE studies containing Type I reliability induction, none were empirically defensible.

Table 6

Alcohol Screening Measures: Reliability Reporting Practices.

Measure	Published Studies	Reported	RI: Type I	RI: Type II
AUDIT	104	22.1%	4.8%	67.3%
MAST	446	11.0%	20.0%	68.6%
CAGE	259	7.3%	5.0%	87.6%

Note. "Published Studies" refers to the number of peer reviewed journal articles that contain the relevant measure and were analyzed in the present research. "Reported" refers to studies that reported score reliability based on the data at hand. "RI: Type I" refers to studies that utilized Type I reliability induction and "RI: Type II" refers to studies that used Type II reliability induction. Further, MAST studies refer to available studies on the MAST, SMAST, and BMAST.

Meier & Davis (1990) noted that giving psychometric information, including reliability estimates, in primary research by citing other studies leads consumers of that research to at least two assumptions: (a) That the citations adequately describe the psychometric properties they reference, and (b) that the properties they reference adequately meet available standards. The present study suggests that consumers of

alcohol screening research generally and CAGE research in particular should use extreme caution when making these assumptions.

The paucity of reliability reporting in the alcohol screening research is striking yet adds to a growing body of literature documenting the under-reporting of reliability information in the empirical literature more generally and lends credence to the decision by scientific journals to offer reliability reporting guidelines to authors submitting papers (e.g., *Educational and Psychological Measurement*).

This observation is unfortunate given how broadly these measures have been employed. They have been used in more than 20 countries as well as with a variety of ethnic groups within the United States. The AUDIT, CAGE, MAST, and SMAST have been employed to evaluate a number of high-risk groups including the elderly, pregnant woman, and emergency room, trauma, triage, general hospital, psychiatric, and family practice patients. Despite their use across remarkably broad samples, we still lack the data necessary to make even the most basic distinctions about how reliably their scores perform among even broad sample characteristics (e.g., gender or age). The fact that almost 90% of the reviewed publications failed to present reliability information, and therefore could not be included in the present analyses, suggests that results from the present studies must be interpreted with caution.

Reliability Generalization

Although the reliability generalization applications are aimed at characterizing score reliability among measures (see Table 7), it is important to note that they are *not* capable of demonstrating whether a test is reliable or unreliable. Reliability is a function of scores and not tests and results from the present application should not be interpreted

to mean that the MAST is more reliable than the AUDIT and the SMAST and that the AUDIT and the SMAST are more reliable than the CAGE. What can be concluded from Table 7, however, is that the MAST tends to generate more reliable *scores* than the AUDIT and the SMAST and the AUDIT and the SMAST tend to generate more reliable *scores* than the CAGE.

Table 7

Alcohol Screening Measures: Descriptive Statistics for Score Reliability.

Measure	Mean	Median	SD	Range	<i>n</i>
MAST	.82	.85	.10	.47 – .91	33
AUDIT	.79	.81	.10	.59 – .94	24
SMAST	.79	.84	.10	.57 – .93	21
CAGE	.74	.74	.09	.52 – .90	22

By comparing the results presented in Table 7 to current standards of acceptable reliability coefficient levels, conclusions can be drawn about the proper use of the alcohol screening measures reviewed here. While only crude estimates, current standards suggest a minimum score reliability cut-off value of .70 for the early stage of measure development, .80 for basic research purposes, and .90 when important clinical decisions are being made (Nunnally & Bernstein, 1994). The AUDIT, CAGE, MAST, and SMAST are not measures in the early stage of development and are widely used for basic research purposes, therefore, score reliability should, *at least*, be at or above .80. While the mean and median values of the AUDIT, MAST, and SMAST score reliabilities

indicate adequate proportions of true score variance for basic research purposes, the CAGE values do not. Because AUDIT, CAGE, MAST, and SMAST score reliability estimates fall short of the recommended minimum cut-off of .90, individuals using these measures, especially in high-stakes situations, should do so with caution. In situations where personal and social costs are significant (e.g., diagnosis), interpreting AUDIT, CAGE, MAST, and SMAST scores in isolation should be avoided altogether.

In addition, reliability generalization studies offer an important opportunity to explore factors associated with variations in score reliability. Results from the AUDIT and MAST applications converge with classical test theory (Lord & Novick, 1968) and clearly indicated that score variability was strongly related to score reliability. If larger sample size had been obtained for the CAGE and SMAST applications, similar results would have been expected. Therefore, researchers employing these AUDIT, CAGE, MAST, and SMAST should take steps toward obtaining sufficiently heterogeneous respondent samples if they are interested in optimizing the performance of the measures.

An important trend also emerged in the relationship between the score reliability of the evaluated measure and sample age. In the AUDIT application, there was a negative and statistically significant relationship between score reliability and age indicating that younger respondents tend to produce more reliable scores than older ones. In the CAGE and MAST applications, there was a positive and statistically significant relationship between score reliability and age indicating that older respondents tend to produce more reliable scores than younger ones. Therefore, the trend toward replacing cumbersome adolescent alcohol screening measures with the brief assessments evaluated here (e.g., Chung et al., 2000; Thom et al., 1999) should be done with caution.

The AUDIT, CAGE, MAST, and SMAST are “brief alcohol screening measures” and have been fittingly praised for their succinctness and ease of administration. Further, it is appropriately noted that their low number of items likely restricts their respective score reliability estimates. Nevertheless, alcohol problems are a major contemporary public health concern and they come at overwhelming personal, social, and economic costs. Furthermore, the vast majority of people with alcohol problems will go untreated (Ogborne & DeWit, 1999; Wu, Kouzis, & Leaf, 1999). Because the misidentification and under-identification of alcohol disorders can contribute to these problems, the time saved administering shorter screening instruments may not be worth the long-term costs. Researchers and clinicians alike recognize this point and have taken steps toward using brief measures in conjunction with other questions. For example, in the *Physicians Guide to Helping Patients with Alcohol Problems* (NIAAA, 1995b), the CAGE is combined with three questions on quantity, frequency, and maximum consumption. In light of the present results, an endeavor such as this should be praised and encouraged.

Limitations

There are several limitations relevant to the reliability generalizations in the current project. First, as noted, poor reliability reporting in the AUDIT, CAGE, MAST, and SMAST literature led to relatively small sample size per application. Small sample size, consequently, compromises the integrity of the results and, therefore, they must be held as tentative until more reliability coefficients become available. The fact that only 22.1% of AUDIT studies, 7.3% of CAGE studies, and 11% of MAST/SMAST/BMAST studies provided reliability information for the data at hand is disappointing. This adds to a growing body of literature demonstrating inadequacy of reliability reporting across a

wide variety of assessment domains (cf., Vach-Haase et al., in press). This is perhaps reflective of a lack of understanding of score reliability or a lack of awareness of current reporting standards. Reading and following the guidelines for authors of Thompson (1994) and the recommendations of Wilkinson and the APA/TFSI (1999) is strongly encouraged. Indeed, the Task Force recommended that authors “provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric” (p. 596), because “Interpreting the size of observed effects requires an assessment of the reliability of the scores” (p. 596).

In addition to poor score reliability reporting, this research uncovered deficient psychometric and descriptive data reporting in the alcohol screening research more generally. This also converges with previous research demonstrating that basic descriptive statistics, above and beyond score reliability, are not always readily available in the psychological literature (Kieffer et al., 2001; Meier & Davis, 1990). In the present applications, the only required pieces of information necessary for study inclusion were score reliability, score variability, average sample age, gender and ethnic breakdown of the sample, and whether the sample consisted of a patient group or not. Using the CAGE literature as an example, only 6 of the 259 publications simultaneously provided even this minimal amount of information. Failure to report score reliability and other important psychometric information (e.g., effect sizes) inhibits scientific progress generally (cf., Thompson, 2002). However, the effects are also local as it resulted in the use of less sophisticated and powerful statistical procedures in the present CAGE, MAST, and SMAST applications. These poor reporting practices also suggests that alcohol assessment researchers still have substantial room for improvement in terms of the

inclusion of complete psychometric data for the scores collected in their studies. Journal reviewers and editors are in a particularly efficacious position from which to encourage, or even require (see Thompson, 1994), complete data reporting.

Several other challenges exist in terms of interpreting the present results. Qualitative and quantitative research summaries may be influenced by biases existing within the population of published studies examined (e.g., McNemar, 1960). There are at least two types of bias that may influence the results of the present applications. The first bias, the “file-drawer” problem (Rosenthal, 1979), exists in virtually all meta-analyses. File-drawer studies are studies considered unacceptable for publication and, therefore, remain in the file-drawers of researchers. This may be so for a variety of reasons, including a general intolerance of poor score reliability and the attenuation of effect sizes and statistical power that accompany low score reliability (Henson, 2001; Thompson, 1994). Second, because of the intolerance for low score reliability, researchers who examine the reliability of their scores and find it low may not report it or may instead report reliability coefficients via induction. Therefore, the reliability coefficients examined in the present applications are likely higher than those that would be found if all research employing the AUDIT, CAGE, MAST, and SMAST had been obtained. This does not bode well for the CAGE questionnaire in particular. In other words, if unpublished reports were obtained, they would likely decrease an already marginally acceptable index of score reliability in CAGE scores.

Summary

The present study introduces the method of reliability generalization into the larger alcohol screening literature and is likely the first reliability generalization

performed in the realm of addictive behaviors research and measurement. Specifically, this study provides a comprehensive examination of three commonly used alcohol screening measures and evaluates the typical reliability of scores that they provide, variations in reliability of scores in sub-populations, and the sample and methodological predictors of variation in score reliability. Results from the present analyses suggest wide variability in the reliability of scores between different alcohol screening measures and within single measures among different populations. This information can help both practitioners and researchers decide which instruments to use with which populations and in which settings.

In addition, the present study required an extensive literature review that gives a framework for continued accumulation and dissemination of empirically supported information regarding the reliability of alcohol screening measures. Reporting these results can raise awareness among alcohol researchers on the necessity of including complete psychometric and demographic information in all studies. In this way, gaps in the understanding and utilization of such screening measures can be filled. *Healthy People 2000: National Health and Disease Objectives for Reducing Alcohol and Other Drug Problems* identified increased utilization of alcohol screening measures as a primary goal for the years 2000 and beyond (Department of Health and Human Services, 1990). By generating increased awareness of, and empirical guidelines for, the accurate use of alcohol screening measures, this research assists in meeting that goal.

References

References marked with a number one (¹) indicate studies included in the AUDIT meta-analysis.

References marked with a number two (²) indicate studies included in the CAGE meta-analysis.

References marked with a number three (³) indicate studies included in the MAST meta-analysis.

References marked with a number three (⁴) indicate studies included in the SMAST meta-analysis.

Adams, W. L., Magruder-Habib, K., Trued, S., & Broome, H. L., (1992). Alcohol abuse in elderly emergency department patients. *Journal of the American Geriatric Society*, 40, 1236-1240.

Adams, W. L., Yuan, Z., Barboriak, J. J., & Rimm, A. A. (1993). Alcohol related hospitalizations of elderly people: Prevalence and geographic variation in the United States. *Journal of the American Medical Association*, 270, 1222-1225.

⁴Al-Ansari, E. A., & Negrete, J. C. (1990). Screening for alcoholism among alcohol users in a traditional Arab Muslim society. *Acta Psychiatrica Scandinavica*, 81, 284-288.

⁴Alexander, D., & Mangelsdorff, A. D. (1994). Screening alcohol abuse potential among Army reservists with the Short Michigan Alcoholism Screening Test. *Military Medicine*, 159, 595-600.

Allen, J. P., Cross, G. M., Fertig, J. B., & Litten, R. Z. (1998). Screening for alcohol problems in the military: Recommended tests. *Military Medicine*, 163, 9-12.

Allen, J. P., & Fertig, J. B. (1995). Assessment in alcoholism treatment: An overview. In J. P. Allen & M. Columbus (Eds.), *Assessing Alcohol Problems: A Guide for Clinicians and Researchers*. National Institute of Health: Bethesda, MD. National Institute of Health, NIH No. 95-3745.

Barnes, L. L. B., Harp, D., & Jung, W. S. (in press). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*.

^{1,4}Barry, K. L., & Fleming, M. F. (1993). The alcohol use disorders identification test (AUDIT) and the SMAST-13: Predictive validity in a rural primary care sample. *Alcohol & Alcoholism*, 28, 33-42.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (in press). A reliability generalization study of the Marlow-Crowne Social Desirability Scale. *Educational and Psychological Measurement*.

³Boyd, M. R., & Hauenstein, E. J. (1997). Psychiatric assessment and confirmation of dual disorders in rural substance abusing women. *Archives of Psychiatric Nursing*, 11, 74-81.

Bradley, K. A., Boyd-Wickizer, J., Powell, S. H., & Burman, M. L. (1998). Alcohol screening questionnaires in women. *Journal of the American Medical Association*, 280, 166-171.

Bradley, K. A., Kivlahan, D. R., Bush, K. R., McDonell, M. B., & Fihn, S. D. (2001). Variations on the CAGE alcohol screening questionnaire: Strengths and limitations in VA general medical patients. *Alcoholism: Clinical and Experimental Research*, 25, 1472-1478.

Bush, B., Shaw, S., Cleary, P., Delbanco, T., & Aronson, M. (1987). Screening for alcohol abuse using the CAGE questionnaire. *The American Journal of Medicine*, 82, 231-5.

- Capraro, R. M., & Capraro, M. M. (in press). Myers-Briggs Type Indicator score reliability across studies: A meta-analytic reliability generalization study. *Educational and Psychological Measurement*.
- Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the mathematics anxiety rating scale across studies. *Educational and Psychological Measurement*, 61, 373-386.
- Carey, K. B. & Teitelbaum, L. M. (1996). Goals and methods of alcohol assessment. *Professional Psychology: Research and Practice*, 27, 460-466.
- Caruso, J. C. (2000). Reliability generalization of the NEO Personality Scales. *Educational and Psychological Measurement*, 60, 236-254.
- Caruso, J. C., & Edwards, S. (2001). Reliability generalization of the Junior Eysenck Personality Questionnaire. *Personality and Individual Differences*, 31, 173-184.
- Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J. D. (2001). Reliability of Scores from the Eysenck Personality Questionnaire: A Reliability Generalization Study. *Educational and Psychological Measurement*, 61, 675-689.
- ³Chalmers, D., Olenick, N. L., & Stein, W. (1993). Dispositional traits as risk in problem drinking. *Journal of Substance Abuse*, 5, 401-410.
- Cherpitel, C. J. (1998). Differences in performance of screening instruments for problem drinking among Blacks, Whites, and Hispanics in an emergency room population. *Journal of Studies on Alcohol*, 59, 420-426.

- Cherpitel, C. J. & Borges, G. (2000). Screening instruments of alcohol problems: A comparison of cut points between Mexican American and Mexican Patients in the emergency room. *Substance Use and Misuse*, 35, 1419-1430.
- Cherpitel, C. J. & Clark, W. B. (1995). Ethnic differences in performance of screening instruments for identifying harmful drinking and alcohol dependence in the emergency room. *Alcoholism: Clinical and Experimental Research*, 19, 628-634.
- Chung, T., Colby, S. M., Barnett, N. P., Rohsenow, D. J., Spirito, A., & Monti, P. M. (2000). Screening adolescents for problem drinking: Performance of brief screens against DSM-IV alcohol diagnoses. *Journal of Studies on Alcohol*, 61, 579-587.
- Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, 21, 360-375.
- ^{2,3}Clements, R. (1998). A critical evaluation of several alcohol screening instruments using the CIDI-SAM as a criterion measure. *Alcoholism: Clinical and Experimental Research*, 22, 985-993.
- Conigrave, K., Hall, W. D., & Saunders, J. B. (1995). The AUDIT questionnaire: Choosing a cut-off score. *Addiction*, 90, 1349-1356.
- Connors, R. J., O'Farrell, T. J., Cutter, H. S. G., & Logan, Thompson, D. L. (1987). Dose-related effects of alcohol among male alcoholics, problems drinkers, and nonproblem drinkers. *Journal of Studies on Alcohol*, 48, 461-466.
- ³Conley, T. B. (2001). Construct validity of the MAST and AUDIT with multiple offender drunk drivers. *Journal of Substance Abuse Treatment*, 20, 287-295.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin.

- ^{2,4}Cornel, M., Knibbe, R. A., van Zutphen, W. M., & Drop, M. J. (1994). Problem drinking in a general practice population: The construction of an interval scale for severity of problem drinking. *Journal of Studies on Alcohol*, 55, 466-70.
- ³Crook, G. M., Oei, T. P., & Young, R. M. (1994). Structure of the MAST with an Australian sample of alcoholics. *Drug & Alcohol Review*, 13, 41-46.
- ¹Daepfen, J-B., Yersin, B., Landry, U., Pecoud, A., & Decrey, H. (2000). Reliability and validity of the Alcohol Use Disorders Identification Test (AUDIT) imbedded within a general health risk screening questionnaire: Results of a survey in 332 primary care patients. *Alcoholism: Clinical and Experimental Research*, 24, 659-665.
- ¹Davey, J. D., Obst, P. L., & Sheehan, M. C. (2000). The use of the AUDIT as a screening tool for alcohol use in the police work-place. *Drug and Alcohol Review*, 19, 49-54.
- ¹Dawe, S., Seinen, A., & Kavanagh, D. (2000). An examination of the utility of the AUDIT in people with schizophrenia. *Journal of Studies on Alcohol*, 61, 744-750.
- Deditius-Island, H. K., & Caruso, J.C. (In press). An examination of the reliability of scores from Zuckerman's sensation seeking scales, form V. *Educational and Psychological Measurement*.
- Department of Health and Human Services (1990). *Healthy People 2000: National Health Promotion and Disease Prevention Objectives*. Rockville, MD: US Department of Health and Human Services, Public Health Services.

- Dimitrov, D. M. (In press). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*.
- Duszynski, K. R., Nicto, F. J., & Valente, C. M. (1995). Repeated practices, attitudes, and confidence levels of primary care physicians regarding patients who abuse alcohol and other drugs. *Maryland Medical Journal*, 44, 439-446.
- ^{2, 4}Dyson, V., Appleby, L., Altman, E., Doot, M., Luchins, D. J., & Delehant, M. (1998). Efficiency and validity of commonly used substance abuse screening instruments in public psychiatric patients. *Journal of Addictive Diseases*, 17, 57-77.
- Ewing, J. A. (1984). Detecting alcoholism: The CAGE questionnaire. *Journal of the American Medical Association*, 252, 1905-1907.
- Fiellin, D. A., Reid, C., & O'Connor, P. G. (2000). Screening for alcohol problems in primary care: A systematic review. *Archives of Internal Medicine*, 160, 1977-1989.
- Fleming, M. F. (1997). Strategies to increase alcohol screening in health care settings. *Alcohol, Health, & Research World*, 21, 340-347.
- ⁴Fleming, M. F., & Barry, K. L. (1989). A study examining the psychometric properties of the SMAST-13. *Journal of Substance Abuse*, 1, 173-182.
- ^{2, 4}Fleming, M., & Barry, K. (1991). A three-sample test of a masked alcohol screening questionnaire. *Alcohol and Alcoholism*, 26, 81-91.
- ¹Fleming, M. F., Barry, K. L., & MacDonald, R. (1991). The alcohol use disorders identification test (AUDIT) in a college sample. *The International Journal of the Addictions*, 26, 1173-1185.

- Gallanter, M., Castaneda, R., & Ferman, J. (1988). Substance abuse among general psychiatric patients. : Place of presentation, diagnosis, and treatment. *American Journal of Drug and Alcohol Abuse*, 14, 211-235.
- Grimm, L. G., & Yarnold, P. R. (1995). *Reading and Understanding Multivariate Statistics*. Washington D.C.: American Psychological Association.
- Grimm, L. G., & Yarnold, P. R. (2000). *Reading and Understanding More Multivariate Statistics*. Washington D.C.: American Psychological Association.
- Hanson, W. E., Curry, K. T., & Bandalos, D. L. (in press). Reliability generalization of Working Alliance Inventory scale scores. *Educational and Psychological Measurement*.
- ^{2, 4}Hays, R. D., Hill, L., Gillogly, J. J., Lewis, M.W., Bell, R. M., Nicholas, R. (1993). Response times for the CAGE, Short-MAST, AUDIT, and JELLINEK alcohol scales. *Behavior Research Methods, Instruments & Computers*, 25, 304-307.
- ^{1, 2, 4}Hays, R. D., & Merz, J. F., & Nicolas, R. (1995). Response burden, reliability, and validity of the CAGE, short MAST, and AUDIT alcohol screening measures. *Behavior Research Methods, Instruments, & Computers*, 27, 277-280.
- ⁴Hays, R. D., & Revetto, J. P. (1992). Old and new MMPI-derived scales and the short-MAST as screening tools for alcohol disorder. *Alcohol & Alcoholism*, 27, 685-695.
- Heck, E. J. & Lichtenberg, J. W. (1990). Validity of the CAGE in screening for problem drinking in college students. *Journal of College Student Development*, 31, 359-364.

- Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, 34, 177-189.
- Henson, R. K., & Hwang, D. (in press). Variability and prediction of measurement error in Kolb's Learning Style Inventory scores: *A reliability generalization study. Educational and Psychological Measurement*.
- Henson, R. K., & Thompson, B. (in press). Characterizing measurement error in scores across studies: Some recommendations for conducting "Reliability Generalization" (RG) studies. *Measurement and Evaluation in Counseling and Development*.
- Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and Psychological Measurement*, 61, 404-420.
- ¹Hermansson, U., Helander, A., Huss, A., Brandt, L., & Ronnberg, S. (2000). The alcohol use disorders identification test (AUSIT) and carbohydrate-deficient transferrin (CDT) in a routine workplace health examination. *Alcoholism: Clinical and Experimental Research*, 24, 180-187.
- Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, 60, 523-531.
- Holder, H., Longabaugh, R., Miller, W.R., & Rubonis (1991). The cost-effectiveness of treatment for alcoholism: A first approximation. *Journal of Studies on Alcohol*, 52, 517-540.

- ³Hotch, D. F., Sherin, K. M., Harding P. N, & Zitter, R. E. (1983). Use of the self-administered Michigan Alcoholism Screening Test in a family practice center. *Journal of Family Practice*, 17, 1021-1026.
- Hunter, J. E., & Schmidt, F. L. (1990). *Method of Meta-Analysis*. Newbury Park, CA: Sage.
- ¹Ivis, F. J., Adlaf, E. M., & Rehm, J. (2000). Incorporating the AUDIT into a general population telephone survey: a methodological experiment. *Drug and Alcohol Dependence*, 60, 97-104.
- Joseph, C. L., Rasmussen, J., Ganzini, L., & Atkinson, R. M. (1997). Outcome of nursing home care for residents with alcohol use disorders. *International Journal of Geriatric Psychiatry*, 12, 767-772.
- ³Kannappan, R. (1988). Screening test and alcoholism. *Indian Journal of Psychological Medicine*, 11, 7-10.
- ¹Karno, M., Granholm, E. & Lin, A. (2000). Factor structure of the alcohol use disorders identification test (AUDIT) in a mental health clinic sample. *Journal of Studies on Alcohol*, 61, 751-758.
- Kazdin, A. E. (1992). *Research Design in Clinical Psychology*. Boston: Allyn & Bacon.
- Kieffer, K. M., & Reese, R. J. (in press). A reliability generalization study of the Geriatric Depression Scale (GDS). *Educational and Psychological Measurement*.
- Kieffer, K M., Reese, R. J., Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *The Journal of Experimental Review*, 69, 280-309.

- Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook* (3rd ed.). New Jersey: Prentice Hall.
- Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H-U., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of the DSM-III-R psychiatric disorders in the United States. *Archives of General Psychiatry*, 51, 8-19.
- ³Knight, B. G., & Mjeldre-Mossey, L. A. (1995). A comparison of the Michigan Alcoholism Screening Test and the Michigan Alcoholism Screening Test-Geriatric Version in screening for higher alcohol use among dementia caregivers. *Journal of Mental Health & Aging*, 1, 147-155.
- Lane, G. G., White, A. E., & Henson, R. K. (in press). Expanding Reliability Generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement*.
- ³Langevin, R., & Lang, R. A. (1990). Substance abuse among sex offenders. *Annals of Sex Research*, 3, 397-424.
- ¹Lapham, S. C., Brown, P., Suriyawongpalsal, P., Skipper, B. J., Chadbimchachai, W., Palsarnslip, S. (1999). Use of the AUDIT for alcohol screening among emergency room patients in Thailand. *Substance Use and Misuse*, 34, 1881-1895.
- ¹Lennings, C. J. (1999). An evaluation of the Leeds Dependence Questionnaire. *Journal of Child & Adolescent Substance Abuse*, 8, 73-87.

- ¹Lennings, C. J., Feeney, G. F., Sheehan, M., Young, R., McPherson, A., & Tucker, J. (1997). Work-place screening of mine employees using the alcohol use disorders identification test (AUDIT) and alcohol breathalyzation. *Drug and Alcohol Review, 16*, 357-363.
- ¹Lennings, C. J., Scott, L., Harris, J., Kingsbury, A., Andrew, M. (1998). Dual diagnosis (comorbidity) among clients of a drug and alcohol outpatient service. *Alcoholism Treatment Quarterly, 16*, 79-87.
- Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1*, 98-107.
- Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Maisto, S. A., Carey, M. P., Carey, K. B., Gordon, C. M., & Gleason, J. R. (2000). Use of the AUDIT and the DAST-10 to identify alcohol and drug use disorders among adults with a severe and persistent mental illness. *Psychological Assessment, 12*, 186-192.
- Maisto, S. A., Conigliaro, J., McNeil, M., Kraemer, K., & Kelley, E. (2000). An empirical investigation of the factor structure of the AUDIT. *Psychological Assessment, 12*, 346-353.
- Maisto, S. A., Connors, G. J., & Allen, J. P. (1995). Contrasting self-report screens for alcohol problems: A review. *Alcoholism: Clinical and Experimental Research, 24*, 1510-1516.

- Mayfield, D., McLeod, G., & Hall, P. (1974). The CAGE questionnaire: Validation of a new alcoholism instrument. *American Journal of Psychiatry*, 131, 1121-1123.
- ¹McCann, B. S., Simpson, T. L., Ries, R., Roy-Byrne, P. (2000). Reliability and validity of screening instruments for drug and alcohol abuse in adults seeking evaluation for attention-deficit/hyperactivity disorder. *The American Journal on Addictions*, 9, 1-9.
- McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist*, 15, 295-300.
- McPherson, T. L., & Hersch, R. K. (2000). Brief substance use screening instruments for primary care settings: A review. *Journal of Substance Abuse Treatment*, 18, 193-202.
- Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology*, 37, 113-115.
- Messick, S., & the Educational Testing Service (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- ²Midanik, L. T., Zahnd, E. G., & Klein, D. (1998). Alcohol and drug CAGE screeners for pregnant low-income women: The California perinatal needs assessment. *Alcoholism: Clinical & Experimental Research*, 22, 121-125.
- ^{2,3}Mischke, H. D., & Venneri, R. L. (1987). Reliability and validity of the MAST, Mortimer-Filkins Questionnaire and CAGE in DWI assessment. *Journal of Studies on Alcohol*, 48, 492-501.

- Moore, R. D., Bone, L. R., Geller, G., Mamon, J. A., Stokes, E. J., & Levine, D. M. (1989). Prevalence, detection, and treatment of alcoholism in hospitalized patients. *Journal of the American Medical Association*, 261, 403-407.
- ²Murray, M., & McMillian, C. (1993). Problem drinking in Northern Ireland: Results of a community survey using the CAGE questionnaire. *Alcohol & Alcoholism*, 28, 477-483.
- ^{2,3}Myerholtz, L. E., & Rosenberg, H. (1997). Screening DUI offenders for alcohol problems Psychometric assessment of the Substance Abuse Subtle Screening Inventory. *Psychology of Addictive Behaviors*, 11, 155-165.
- ^{2,3}Myerholtz, L. E., & Rosenberg, H. (1998). Screening college students for alcohol problems: Psychometric assessment of the SASSI-2. *Journal of Studies on Alcohol*, 59, 439-446.
- ³Nanakorn, S., Fukuda, K., Nishiyori, A., Shibata, A., & Nakamura, J. (1999). Aldehyde dehydrogenase genotypes and male alcohol use disorders: A case-control study in Khon Kaen, north-east Thailand. *Psychiatry and Clinical Neurosciences*, 53, 397-405.
- National Institute on Alcohol Abuse and Alcoholism (1995a). *Assessing Alcohol Problems: A Guide for Clinicians and Researchers*. National Institute of Health: Bethesda, MD. National Institute of Health, NIH No. 95-3745.
- National Institute on Alcohol Abuse and Alcoholism (1995b). *The Physicians Guide to Helping Patients with Alcohol Problems*. National Institute of Health: Rockville, MD. Department of Health and Human Services No. 95-3769.

- National Institute on Drug Abuse & National Institute on Alcohol Abuse and Alcoholism (1998). The economic costs of alcohol and drug abuse in the United States-1992.
- Nilssen, O., & Cone, H. (1994). Screening patients for alcohol problems in primary health care settings. *Alcohol Health and Research World*, 18, 136-139.
- Nilsson, J.E., Schmidt, C. K., & Meek, W. D. (in press). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement*.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Nystrom, M., Perasalo, J., & Salaspuro, M. (1993). Screening for heavy drinking and alcohol related problems in young university students: the CAGE, the MmMAST, and the trauma score questionnaire. *Journal of Studies on Alcohol*, 54, 528-533.
- ¹O'Hare, T., & Sheerer, M. V. (1999). Validating the alcohol use disorder identification test with college first-offenders. *Journal of Substance Abuse Treatment*, 17, 113-119.
- ²O'Hare, T., & Tran, T. (1997). Predicting problem drinking in college students: Gender differences and the CAGE questionnaire. *Addictive Behaviors*, 22, 13-21.
- Ogborne, A. C. & DeWit, D. J. (1999). Lifetime use of professional and community services for help with drinking: Results from a Canadian population survey. *Journal of Studies on Alcohol*, 60, 867-872.

- Peter, J. P., & Churchill, G. A. (1986). Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. *Journal of Marketing Research*, 23, 1-10.
- Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21, 381-392.
- Piccinelli, M., Tessari, E., Bortolomasi, Piasere, O., Semenzin, M., Garzotto, N., & Tansella, M. (1997). Efficacy of the alcohol use disorders identification test as a screening tool for hazardous alcohol intake and related disorders in primary care: A validity study. *British Medical Journal*, 314, 420-424.
- Pokorny, A. D., Miller, B. A., & Kaplan, H. B. (1972). The brief MAST: A shortened version of the Michigan Alcoholism Screening Test. *American Journal of Psychiatry*, 129, 342-345.
- Reese, R. J., Kieffer, K. M., & Briggs, B. K. (in press). A reliability generalization study of select measures of adult attachment style. *Educational and Psychological Measurement*.
- Reinherz, H. Z., Giaconia, R. M., Lefkowitz, E. S., & Pakiz, B. (1993). Prevalence of psychiatric disorders in a community population of older adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 32, 369-377.
- Reynaud, M., Malet, L., Facy, F., & Glanddier, P. (2000, Jul). Hospital morbidity of alcohol use disorders in the center of France. *Alcoholism: Clinical and Experimental Research*, 24, 1057-62.

- ³Richman, J. A., Flaherty, J. A., & Pyskoty, C. (1992). Shifts in problem drinking during a life transition: Adaptation to medical school training. *Journal of Studies on Alcohol*, 53, 17-24.
- ³Rodney, H. E. (1995). A profile of collegiate Black adult children of alcoholics. *Journal of College Student Development*, 36, 228-235.
- ³Rodney, H. E., & Rodney, L. (1996). An exploratory study of African American collegiate adult children of alcoholics. *Journal of American College Health*, 44, 267-72.
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Ross, H., Gavin, D., & Skinner, H. (1990). Diagnostic validity of the MAST and the alcohol dependence scale in the assessment of DSM-II alcohol disorders. *Journal of Studies on Alcohol*, 51, 506-513.
- ^{2,3,4}Rumpf, H. J., Hapke, U., & Hill, John, U. (1997). Development of a screening questionnaire for the general hospital and general practices. *Alcoholism: Clinical & Experimental Research*, 21, 894-898.
- Rydon, P., Redman, S., Sanson-Fisher, R. W., & Reid, A. L. (1992). Detection of alcohol-related problems in general practice. *Journal of Studies on Alcohol*, 53, 197-202.
- ³Saltstone, R., Halliwell, S., & Hayslip M. A. (1994). A multivariate evaluation of the Michigan Alcoholism Screening Test and the Drug Abuse Screening Test in a female offender population. *Addictive Behaviors*, 19, 455-462.

- Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R. & Grant, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction*, 88, 791-803.
- Schuerger, J. M., Tait, E., & Tavernelli, M. (1982). Temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology*, 43, 176-182.
- Schuerger, J. M., Zarrella, K. L., & Hotz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire, *Journal of Personality and Social Psychology*, 56, 777-783.
- Selzer, M. L. (1971). The Michigan Alcoholism Screening Test: The quest for a new diagnostic instrument. *American Journal of Psychiatry*, 127, 1653-1658.
- ^{3, 4}Selzer, M. L., Vinokur, A., Van Rooijan, L. (1975). A self-administered Short Michigan Alcoholism Screening Test (SMAST). *Journal of Studies on Alcohol*, 36, 117-126.
- ³Sheridan, D. P., Johnson, N. P., Michels, P. J., Marion, L. N., & Fuller, S. G. (1995). Substance abuse in freshmen medical students over time. *Substance Abuse*, 16, 93-98.
- Shields, A. L., & Caruso, J. C. (in press). Reliability Generalization of the Alcohol Use Disorders Identification Test. *Educational and Psychological Measurement*.
- Shields, A. L., & Caruso, J. C. (under review). A reliability induction and reliability generalization of the CAGE questionnaire.

Shields, A. L., Guttmanova, K. & Caruso, J. C. (under review). An examination of the factor structure of the alcohol use disorders identification test in two high-risk samples.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529-540.

³Skinner, H. A. (1979). A multivariate evaluation of the MAST. *Journal of Studies on Alcohol*, 40, 831-844.

³Skinner, H. A., & Allen, B. A. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting & Clinical Psychology*, 51, 267-275.

³Skinner, H. A., & Sheu, W. (1982). Reliability of alcohol use indices: The Lifetime Drinking History and the MAST. *Journal of Studies on Alcohol*, 43, 1157-1170.

¹Skipsey, K., Burleson, J. A. & Kranzler, H. R. (1997). Utility of the AUDIT for identification of hazardous or harmful drinking in drug-dependent patients. *Drug and Alcohol Dependence*, 45, 157-163.

Smith, D. S., Collins, M. Kreisberg, J. P. Volpicelli, J. R., & Alterman, A. (1987). Screening for problem drinking in college freshman. *Journal of American College Health*, 36, 89-94.

Smith-Black, B. Rabins, P. V., & McGuire, M. H. (1998). Alcohol and mortality: Alcohol use disorder is a risk factor for mortality among older public housing residents. *International Psychogeriatrics*, 10, 309-327.

Soderstrom, C. A., Dischinger, P. C., Kerns, T. J., Kufera, J. A., McDuff, D. R., Gorelick, D. A., & Smith, G. S. (1998). Screening trauma patients for alcoholism

- according to NIAAA guidelines with alcohol use disorders identification test questions. *Alcoholism: Clinical and Experimental Research*, 22, 1470-1475.
- Steinbauer, J. R., Cantor, S. B., Holzer, C. E., & Volk (1998). Ethnic and sex bias in primary care screening tests for alcohol use disorders. *Annals of Internal Medicine*, 129, 353-362.
- Steinweg, D. L., & Worth, H. (1993). Alcoholism: The keys for the CAGE. *The American Journal of Medicine*, 94, 520-523.
- Storgaard, H., Nielsen, S. D., & Gluud, C. (1994). The validity of the Michigan Alcoholism Screening Test (MAST). *Alcohol and Alcoholism*, 29, 493-502.
- Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.) *Reading and understanding more multivariate statistics* (pp 23 – 66). Washington D. C: American Psychological Association.
- ⁴Susser, E., Struening, E. L., & Conover, S. (1989). Psychiatric problems in homeless men: Lifetime psychosis, substance use, and current distress in new arrivals at New York City shelters. *Archives of General Psychiatry*, 46, 845-850.
- Thom, B., Herring, R., & Judd, A. (1999). Identifying alcohol-related harm in young drinkers: The role of accident and emergency departments. *Alcohol and Alcoholism*, 34, 910-915.
- ⁴Thombs, D. L. (1991). Expectancies versus demographics in discriminating between college drinkers: Implications for alcohol abuse prevention. *Health Education Research*, 6, 491-495.
- Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 837-847.

- Thompson, B. (2000). Reliability generalization [Special Section]. *Educational and Psychological Measurement*, 60, 157-254.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effects sizes. *Educational Researcher*, April, 2002.
- Thompson, B. (in press). Reliability generalization [Special Issue]. *Educational and Psychological Measurement*.
- Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *Journal of Counseling & Developmental* research articles. *Journal of Counseling & Development*, 76, 436-441.
- ³Thurber, S., Snow, M., Lewis, D., & Hodgson, J. M. (2001). Item characteristics of the Michigan Alcoholism Screening Test. *Journal of Clinical Psychology*, 57, 139-144.
- Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, 58, 6-20.
- Vacha-Haase, T., Henson, R., & Caruso, J. C. (in press). Reliability Generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*.
- Vacha-Haase, T., Kogan, L. R., Tani, C. R., & Woodall, R. A. (2001). Reliability generalization: Exploring variation of reliability of MMPI clinical scales scores. *Educational and Psychological Measurement*, 61, 45-59.

- Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522.
- Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10*, 413-425.
- Vacha-Haase, T., Tani, C. R., Kogan, L. R., Woodall, R. A., & Thompson, B. (2001). Reliability generalization: Exploring reliability variations on MMPI/MMPI-2 validity scale scores. *Assessment, 8*, 391-401.
- Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*, 224-235.
- Volk, R. J., Steinbauer, J. R., Cantor, S. B., & Holzer, C. E. (1997). The Alcohol Use Disorders Identification Test (AUDIT) as a screen for at-risk drinking in primary care patients of different racial/ethnic backgrounds. *Addiction, 92*, 197-206.
- Wallace, K. A., & Wheeler, A. J. (in press). Reliability generalization of the Life Satisfaction Index. *Educational and Psychological Measurement*.
- Weisner, C., Greenfield, T., & Room, R. (1995). Trends in the treatment of alcohol problems in the US general population, 1979 through 1990. *American Journal of Public Health, 85*, 55-60.
- Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58*, 21-37.

- Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. *Educational Researcher*, 9, 5-10.
- Wilkinson, L., & the American Psychological Association Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist*, 54, 594-604.
- Wolford, G. L., Rosenberg, S. D., Drake, R. E., Mueser, K. T., Oxman, T. E., Hoffman, D., Vidaver, R. M., Juckour, R., & Carrieri, K. L. (1999). Evaluation of methods for detecting substance use disorder in persons with severe mental illness. *Psychology of Addictive Behaviors*, 13, 313-326.
- Wu, L-T., Kouzis, A. C., & Leaf, P. J. (1999). Influence of comorbid alcohol and psychiatric disorders on utilization of mental health services in the national comorbidity survey. *American Journal of Psychiatry*, 156, 1230-1236.
- Yin, P., & Fan, X. (2000). Assessing the reliability of Beck depression inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, 60, 201-223.
- Youngstrom, E. A., & Green, K. W. (in press). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement*.
- ^{3,4}Zung, B. J. (1979). Psychometric properties of the MAST and two briefer versions. *Journal of Studies on Alcohol*, 40, 845-854.
- ³Zung, B. J. (1980). Factor structure of the Michigan Alcoholism Screening Test (MAST) in a psychiatric outpatient population. *Journal of Clinical Psychology*, 36, 1024-1030.

³Zung, B. J. (1982). Evaluation of the Michigan Alcoholism Screening Test (MAST) in assessing lifetime and recent problems. *Journal of Clinical Psychology*, 38, 425-439.

⁴Zung, B. J. (1982). Screening for alcohol problems in psychiatric inpatients. *The Psychiatric Forum*, 11, 32-37.

³Zung, B. J., & Ross, M. (1980). Factor structure of the Michigan Alcoholism Screening Test (MAST) among acutely disturbed psychiatric patients. *Journal of Clinical Psychology*, 36, 806-812.

Appendix A.

The Alcohol Use Disorders Identification Test (AUDIT)¹

<u>Item</u>	<u>Question</u>
-------------	-----------------

- | | |
|----|---|
| 1. | How often do you have a drink containing alcohol?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily |
| 2. | How many drinks containing alcohol do you have on a typical day when you are drinking?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily |
| 3. | How often do you have six or more drinks on one occasion?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily |
| 4. | How often during the last year have you found that you were not able to stop drinking once had started?

(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily |

5. How often during the last year have you failed to do what was normally expected from you because of drinking?
- (0) Never
 - (1) Less than monthly
 - (2) Monthly
 - (3) Weekly
 - (4) Daily or almost daily
6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?
- (0) Never
 - (1) Less than monthly
 - (2) Monthly
 - (3) Weekly
 - (4) Daily or almost daily
7. How often during the last year have you had a feeling of guilt or remorse after drinking?
- (0) Never
 - (1) Less than monthly
 - (2) Monthly
 - (3) Weekly
 - (4) Daily or almost daily
8. How often during the last year have you been unable to remember what happened the night before because you had been drinking?
- (0) Never
 - (1) Less than monthly
 - (2) Monthly
 - (3) Weekly
 - (4) Daily or almost daily
9. Have you or someone else been injured as a result of your drinking?
- (0) No
 - (2) Yes, but not in the last year
 - (4) Yes, during the last year

10. Has a relative or friend, or a doctor or other health care worker been concerned about your drinking or suggested you cut down?
- (0) No
 - (2) Yes, but not in the last year
 - (4) Yes, during the last year

¹Scoring: Total AUDIT score is the sum of the endorsed items.

Appendix B.

The CAGE Questionnaire¹

<u>Item</u>	<u>Yes</u>	<u>No</u>	<u>Question</u>
1.	(1)	(0)	Have you ever felt like you should CUT down on your drinking?
2.	(1)	(0)	Have people ANNOYED you by criticizing your drinking?
3.	(1)	(0)	Have you ever felt bad or GUILTY about your drinking?
4.	(1)	(0)	Have you ever had a drink first thing in the morning to steady your nerves or to get rid of a hangover (EYE-OPENER)?

¹Scoring: Total CAGE score is the sum of endorsed items.

Appendix C.

The Michigan Alcoholism Screening Test¹.

<u>Item</u>	<u>Yes</u>	<u>No</u>	<u>Question</u>
1.	(0)	(0)	Do you enjoy a drink now and then?
2.	(0)	(2)	Do you feel you are a normal drinker (by normal we mean you drink less than or as much as most other people)?
3.	(2)	(0)	Have you ever awakened the morning after some drinking the night before and found you could not remember a part of the evening before?
4.	(1)	(0)	Does your boyfriend/girlfriend, spouse, or parents, ever worry or complain about your drinking?
5.	(0)	(2)	Can you stop drinking without a struggle after a one or two drinks?
6.	(1)	(0)	Do you ever feel guilty about your drinking?
7.	(0)	(2)	Do friends or relatives think you are a normal drinker?
8.	(0)	(2)	Are you able to stop drinking when you want to?
9.	(5)	(0)	Have you ever attended a meeting of Alcoholics Anonymous (AA) for your own drinking?
10.	(1)	(0)	Have you gotten into physical fights when drinking?
11.	(2)	(0)	Has drinking ever caused any problems with you and your significant other?
12.	(2)	(0)	Has your significant other (or other family member) ever gone to anyone for help about your drinking?
13.	(2)	(0)	Have you ever lost friends or girlfriends/boyfriends because of drinking?
14.	(2)	(0)	Have you ever gotten in trouble at work because of drinking?
15.	(2)	(0)	Have you ever lost a job because of drinking?
16.	(2)	(0)	Have you ever neglected your obligations, your family, or your work for two or more days in a row because you were drinking?

17. (1) (0) Do you ever drink before noon?
18. (2) (0) Have you ever been told you have liver trouble? Cirrhosis?
19. (2) (0) Have you ever had delirium tremens (DTs), severe shaking, heard voices or seen things that weren't there after heavy drinking?²
20. (5) (0) Have you ever gone to anyone for help about your drinking?
21. (5) (0) Have you ever been in the hospital because of drinking?
22. (2) (0) Have you ever been a patient in a psychiatric hospital or on a psychiatric ward of a general hospital where drinking was part of the problem?
23. (2) (0) Have you ever been seen at a psychiatric or mental health clinic, or gone to a doctor, social worker, or clergyman for help with an emotional problem in which drinking had played a part?
24. (2) (0) Have you ever been arrested for drunk driving or driving after drinking?³
25. (2) (0) Have you ever been arrested, even for a few hours, because of drunk behavior?

¹Scoring: Total MAST score is the sum of endorsed items.

²5 points are added for each episode of delirium tremens.

³2 points are added for each arrest.

Appendix D.

The Brief Michigan Alcoholism Screening Test (BMAST)¹

<u>Item</u>	<u>Yes</u>	<u>No</u>	<u>Question</u>
1.	(0)	(2)	Do you feel you are a normal drinker?
2.	(0)	(2)	Do friends or relatives think you are a normal drinker?
3.	(5)	(0)	Have you ever attended a meeting of Alcoholics Anonymous (AA)?
4.	(2)	(0)	Have you ever lost friends or girlfriends/boyfriends because of drinking?
5.	(2)	(0)	Have you ever gotten into trouble at work because of drinking?
6.	(2)	(0)	Have you ever neglected your obligations, your family, or your work for two or more days in a row because you were drinking?
7.	(2)	(0)	Have you ever had delirium tremens (DTs), severe shaking, heard voices or seen things that weren't there after heavy drinking?
8.	(5)	(0)	Have you ever gone to anyone for help about your drinking?
9.	(5)	(0)	Have you ever been in a hospital because of drinking?
10.	(2)	(0)	Have you ever been arrested for drunk driving or driving after drinking?

¹Scoring: Total BMAST score is the sum of endorsed items.

Appendix E.

The Short Michigan Alcoholism Screening Test (SMAST)¹

<u>Item</u>	<u>Yes</u>	<u>No</u>	<u>Question</u>
1.	(0)	(2)	Do you feel you are a normal drinker? (By normal we mean you drink less than or as much as most other people.)
2.	(1)	(0)	Does your wife, husband, a parent, or other near relative ever worry or complain about your drinking?
3.	(1)	(0)	Do you ever feel guilty about your drinking?
4.	(2)	(0)	Do friends or relatives think you are a normal drinker?
5.	(0)	(2)	Are you able to stop drinking when you want to?
6.	(5)	(0)	Have you ever attended a meeting of Alcoholics Anonymous (AA)?
7.	(2)	(0)	Has drinking ever created problems between you and your wife, husband, a parent or otherwise near relative?
8.	(2)	(0)	Have you ever gotten into trouble at work or school because of drinking?
9.	(2)	(0)	Have you ever neglected your obligations, your family, or your work for two or more days in a row because you were drinking?
10.	(5)	(0)	Have you ever gone to anyone for help about your drinking? If YES, was this other than Alcoholics Anonymous or a hospital? (If YES, code as YES; if NO, code as NO)
11.	(5)	(0)	Have you ever been in a hospital because of drinking? If YES, was this for (a) detox; (b) alcoholism treatment; (c) alcohol-related injuries or medical problems, e.g., cirrhosis or physical injury incurred while under the influence of alcohol (car accident, fight, etc.).
12.	(2)	(0)	Have you ever been arrested for drunken driving, driving while intoxicated, or driving under the influence of alcoholic beverages?
13.	(2)	(0)	Have you ever been arrested, even for just a few hours, because of other drunken behavior?

¹Scoring: Total MAST score is the sum of endorsed items.